AD_____

MIPR NUMBER 96MM6709

TITLE: Evaluation of Spatial Paradigm for Information Retrieval and Exploration (SPIRE) Technology for Trauma Data Analysis

PRINCIPAL INVESTIGATOR: Sam N. Stevens
Augustin J. Calapristi

CONTRACTING ORGANIZATION: Department of Energy, Richland
Richland, Washington 99352

REPORT DATE: December 1997

TYPE OF REPORT: Final

PREPARED FOR: Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1997	Final (19 Apr 96 - 31 Dec 97)

4. TITLE AND SUBTITLE
Evaluation of Spatial Paradigm for Information Retrieval and Exploration (SPIRE) Technology for Trauma Data Analysis

5. FUNDING NUMBERS
96MM6709

6. AUTHOR(S)
Sam N. Stevens
Augustin J. Calapristi

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Department of Energy, Richland
Richland, Washington 99352

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Commander
U.S. Army Medical Research and Materiel Command
Fort Detrick, Frederick, MD 21702-5012

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200*

The U.S. Army Medical Research and Material Command (MRMC) has commissioned the Pacific Northwest National Laboratory to conduct research in applying visualization technologies in a variety of areas of interest to the army. This work included research in the areas of analysis and visualization of information from trauma databases and technical planning documents, exploration of various visualization concepts for analyzing data from medical sensors, reduction of language to mathematics, and technical advancement of text analysis technologies.

A number of visualization tools and techniques were applied in each area. One of the major technologies used was the Spatial Paradigm for Information Retrieval and Exploration (SPIRE™). SPIRE uses advanced statistics to convert free text into mathematical vectors, which can be analyzed and converted into interactive visualizations. It is designed to produce the analysis and visualizations without requiring the user to have any prior knowledge of the data.

This report provides a high-level overview of the activities for the period. Separate reports have been produced for the major areas of research and are attached as appendixes.

14. SUBJECT TERMS
visualization, text analysis, statistics, medical sensor, language, signal processing, analysis cluster, information analysis, text vector

15. NUMBER OF PAGES
204

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

FOREWORD

Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the U.S. Army.

____ Where copyrighted material is quoted, permission has been obtained to use such material.

____ Where material from documents designated for limited distribution is quoted, permission has been obtained to use the material.

____ Citations of commercial organizations and trade names in this report do not constitute an official Department of Army endorsement or approval of the products or services of these organizations.

____ In conducting research using animals, the investigator(s) adhered to the "Guide for the Care and Use of Laboratory Animals," prepared by the Committee on Care and Use of Laboratory Animals of the Institute of Laboratory Resources, National Research Council (NIH Publication No. 86-23, Revised 1985).

____ For the protection of human subjects, the investigator(s) adhered to policies of applicable Federal Law 45 CFR 46.

____ In conducting research utilizing recombinant DNA technology, the investigator(s) adhered to current guidelines promulgated by the National Institutes of Health.

____ In the conduct of research utilizing recombinant DNA, the investigator(s) adhered to the NIH Guidelines for Research Involving Recombinant DNA Molecules.

____ In the conduct of research involving hazardous organisms, the investigator(s) adhered to the CDC-NIH Guide for Biosafety in Microbiological and Biomedical Laboratories.

C. J. Calapp 12/23/97

PI - Signature Date

Table of Contents

Evaluation of Spatial Paradigm for Information Retrieval and Exploration (SPIRE) Technology for Trauma Data Analysis

Final Report

1.0 Introduction

The U.S. Army Medical Research and Material Command (MRMC) commissioned the Pacific Northwest National Laboratory (PNNL) to conduct research in applying visualization technologies in a variety of areas of interest to the Army. This work included research in the following areas:

- analysis and visualization of information from trauma databases and technical planning documents
- exploration of various visualization concepts for analyzing data from medical sensors
- reduction of language to mathematics
- technical advancement of text analysis technologies.

A number of visualization tools and techniques were applied in each area. One of the major technologies used was the Spatial Paradigm for Information Retrieval and Exploration (SPIRE™). SPIRE uses advanced statistics to convert free text into mathematical vectors, which then can be analyzed and converted into interactive visualizations. It is designed to produce these analysis and visualizations without requiring the user to have any prior knowledge of the data.

This report provides a high-level overview of the activities for the period. Separate reports have been produced for the major areas of research and are attached as appendixes.

2.0 Information Analysis: Trauma Data Analysis

2.1 Purpose and Scope

The purpose of this research was to determine if an advanced text visualization tool could produce meaningful analysis of trauma data that were collected from traffic accident databases. Dr. Howard Champion of the University of Maryland provided the data.

2.2 Approach

SPIRE software was used convert the textual trauma information to visual metaphors for analysis. Previous experience with SPIRE has shown success in analyzing free text documents such as newswire feeds or message traffic. The trauma data presented a special

challenge, because the data contained a much higher degree of structure (i.e., formatting) and were characterized by a sparse use of vocabulary.

The data were processed, visualizations were produced, and the underlying statistical attributes of the data were evaluated. A full discussion of the approach and the visualizations produced are contained in Appendix A, "Trauma Data Analysis."

2.3 Conclusions/Recommendations

SPIRE's text analysis algorithms are tuned to process free text (i.e., prose). The trauma data tended to use the same words in many of the documents and contained a very small vocabulary and sentence fragments. This resulted in poor differentiation among the document set. After some adjustment of the statistical parameters used in the analysis, SPIRE did produce visualizations that reflected limited differentiation among the documents, but it was unable to produce other meaningful relationships. The initial results showed some potential for applying SPIRE to small data sets with limited vocabularies. Lessons learned from the analysis of the trauma data were used to develop techniques for processing of subsequent data sets that were received from MRMC. These techniques are discussed in the following section and in Appendix B, "Visualization Research and Development: Structured and Unstructured Text Analysis."

3.0 Information Analysis: Visualization Research and Development for Structured and Unstructured Text Analysis

3.1 Purpose and Scope

This task was a follow-on to the trauma data analysis discussed in section 2.0. The MRMC provided 34 data files containing an assortment of scientific/technical planning and management information; the full report is contained in Appendix B. These files included Accomplishments by Program Element, Scientific and Technical Objectives, and Thrust Areas. Specific areas of research included

- determining if SPIRE's advanced statistical analysis and graphical presentation techniques are valid for processing structured text
- evaluating approaches to improve the processing and presentation of structured text
- researching approaches to correlating and relating different documents sets.

3.2 Approach

To achieve a greater document sample size, user-defined phrases or keywords were used to identify areas in the planning documents where thematic breaks could be expected. This technique was used to break apart 20 of the 34 files into 800 "sub-documents." (The remaining 14 files consisted of simple lists, which did not provide enough topical content or unique structure to be segmented into smaller files.) In addition, 7 Defense Technology Area Plans (TAPS) dealing with biomedical research were harvested from the Internet and segmented into 20 sub-documents. These were added to the 800 MRMC planning documents.

A file-tagging technique was employed to uniquely identify where the document subsets originated. The tags were used as search terms in SPIRE's query tool. The expectation was that all of the document segments would be analyzed by SPIRE and clustered as though they came from a single source. If overlap in the key themes of the planning and execution documents occurred, they could be expected to cluster together. If a high degree of commonality in the thematic content did not occur, then the documents would most likely cluster according to their source.

3.3 Conclusions/Recommendations

Structured text, like that represented by MRMC, can be processed into useful visualizations through the use of document segmentation and tagging techniques. The consolidated MRMC planning document set provides an example of how SPIRE technology can be used to quickly analyze large numbers of documents and provide new insights regarding relationships among information.

The true value of SPIRE is realized when users can interact with the system to provide fresh insights into their data that may not have been apparent using traditional text analysis approaches. The first step toward reaching this goal has been already completed with the installation of SPIRE at the Walter Reed Army Institute of Research (WRAIR). A second step is to continue to advance the software technology of the SPIRE application so that it can be accessed by a greater number of users. In the case of MRMC, this requires changes to enable the SPIRE software to run on personal computer architectures. A third step is to add new functionality such as data harvesting, automated segmentation of large documents, the ability for users to predefine topics for visualization, and support for structured fields that can be used to create additional subsets. Work has already been started in these areas but needs additional sponsorship to continue.

4.0 Visualization of Medical Sensor Data

4.1 Purpose and Scope

MRMC commissioned PNNL to work with the WRAIR Surgery Division to visualize medical sensor data. The Signal Processing Concept Exploration Task was undertaken as part of a larger project to explore the concept of applying SPIRE visualization technologies to several areas of medicine and was a proof-of-concept task.

We applied the SPIRE analysis paradigm to highly sampled multiple sensor medical data. The data for our analysis were from three patients who had undergone open-heart surgery at WRAIR. Each patient was post-operatively monitored at 1000 Hz for several hours with electrocardiogram (EKG), Direct Arterial Pressure, and Pulse Blood Oximeter sensors, each recording simultaneously. Our analysis of the medical sensor data was based on the definition of an "event" as the activity associated with a particular heartbeat for a particular patient. We used a custom pattern-matching technique to detect distinct heartbeats within the data for each sensor.

Several other activities that were not explicitly outlined in the original statement of scope were undertaken by the Signal Processing Concept Exploration Task. These included

delivery of the SPIRE analysis tool, some custom Perl software scripts, and a custom Medline Abstract data set.

4.2 Approach

In SPIRE, the analysis is based on thematic similarity between documents. Documents thus serve as the core objects within the analysis. A text engine automatically extracts the best word features for the entire document collection. These best word features together with the word frequency counts and word associations are then used to create a numerical signal for each document object. This high-dimensional numerical signal contains the "attributes" for the document objects. These "attributes" are used to cluster the documents and then to project down to a 2-dimensional view called Galaxy, which is based on a Principal Components Analysis (PCA).

The SPIRE text-to-signal generator, which converts words within unstructured text to mathematical signals, does not operate on numerical data such as medical sensor data. Thus, we wrote a custom feature selection engine to extract the features in the neighborhood for each heartbeat. We presented separate Principal Component-Based visualizations for each patient based on a different combination of initial variables.

We processed the time series for each sensor individually to identify the onset time for each heartbeat. The onset times across sensors were then organized by the specific heartbeat to which they referred. These windows around each heartbeat were used to define a neighborhood within which other calculations would be made for a given sensor or across sensors that would ultimately be used to calculate features for our analysis.

The features that we calculated for each heartbeat-to-heartbeat event were predefined for the most part by Dr. Frederick Pearce of WRAIR Surgery Division. Dimensionality reduction was then accomplished on this feature matrix using a PCA. The events (or heartbeats) were then visualized in 2- or 3-space using the PCA scores. A full description of the approach, events, and features are found in the full report on this task in Appendix C, "Visualization of Medical Sensor Data: Signal Processing Concept Exploration."

4.3 Conclusions/Recommendations

Appendix C contains several visualizations of the Principal Components derived from the medical sensor data. The visualizations are intriguing because they differ from patient to patient, with the most different visualization belonging to a patient with a pacemaker. Some artifacts of the visualizations suggest further investigation for certain features / events in the data. However, additional ancillary data associated with each patient was not available to do a qualitative validation of the visualizations effectiveness for providing medical insights to the patient conditions.

Appendix C also contains SPIRE visualizations of a medical corpus comprised of Medline abstracts with publication dates between 1994 and spring 1997. This data set was built for this task. In the summer of 1997, this corpus and the SPIRE source code were installed on "medic1" at the WRAIR Surgery Division.

5.0 Language to Mathematics

5.1 Purpose and Scope

This task was to conduct research into the technologies and scope necessary to reduce natural language into a mathematical form. This task initially covered a broad scope of technologies including linguistic processing, visualization of complex, high-dimensional information spaces, and mathematical and statistical approaches for representing language, and others. Based upon discussions with the client, the foci of the language-processing work moved toward the areas of mathematical structure of language and machine translation, with the long-term goal of applying these technologies to overcome language-based issues in the medical field.

5.2 Approach

This activity began with an investigation of the current state of language-processing research. Based upon the initial research, an organizing perspective that we refer to as the "Language Processing Universe" was produced to help organize the vast language processing literature.

Next, research into the state of the science in standard machine translation technology was conducted. The basis of this description is derived from a seminar / short-course in machine translation at the UCLA.

The knowledge base derived during the discovery phase of the project suggested some fundamental regular/mathematical properties of language. A discussion on how these properties have been exploited to address problems in language processing, such as information retrieval, summarization, disambiguation, and translation, is presented in Appendix D, "Language to Mathematics." Finally, some mathematical and statistical regularities of language are discussed.

Based upon the initial findings of this activity, a research agenda was developed that included recommendation for additional study in selected areas and a suite of experiments to advance the technologies.

5.3 Conclusions/Recommendations

Reports for specific research areas can be found in Appendix D, an overview of natural language processing research and current state of machine translation technology; Appendix E, "Applications of Parallel Corpora in Document Analysis," which discusses the results from an experiment using SPIRE statistics and visualization techniques to creating mappings between documents in different languages; and Appendix F, "References for the Language to Mathematics," a Web page printout containing various language-processing references. The references contained in Appendix F are particularly useful for obtaining an expanded view of language-processing capability.

The proposed research agenda, discussed in Appendix E, suggests a near-term effort to investigate automatic assistance in obtaining data needed for the current design of

machine translation software. Examples of such data are "dictionaries" and grammars for languages. Currently, these are constructed laboriously "by hand." In the longer term, the research agenda is focused on estimating more abstract concepts (such as conflicts and engagements) from data.

Bibliography

Brown PF, PV deSouza, RL Merver, VJ Della Pietra, JC. Lai. 1992. "Class-Based n-gram Models of Natural Language." *Computational Linguistics* 18(4) 467-479.

Carbonell J, Y Yang, R Frederking, RD Brown, Y Geng, and D Lee. 1997. "Translingual Information Retrieval: A Comparative Evaluation." In *Proceedings of Fifteenth International Joint Conference on Artificial Intelligence* (IJCAI-97).

Charniak E. 1995. *Parsing with Context-Free Grammars and Word Statistics*, Technical Report CS-95-28, Department of Computer Science, Brown University.
Charniak E. 1996. *Statistical Language Learning*. MIT Press.

Damashek M. 1994. *Gauging Similarity via n-grams: Text Sorting, Categorization and Retrieval in Any Language*. TR-R53-05-94, National Security Agency.

Damashek M. 1995. "Gauging Similarity with n-Grams: Language-Independent Categorization of Text." *Science* (267) 844-848.

Data Desk Ver. 6.0. Data Description, Inc. Ithica, NY

de Boor C. 1978. "A Practical Guide to Splines." *Applied Mathematical Sciences*, vol. 27. Springer-Verlag, New York.

Devlin KJ. 1997. *Goodbye, Descartes: the End of Logic and the Search for a New Cosmology of the Mind*. John Wiley & Sons, New York.

Dumais ST, TK Landauer, and ML Littman. 1996. "Automatic Cross-Linguistic Information Retrieval Using Latent Semantic Indexing." In *SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval*, pp. 16-23.
http://superbook.bellcore.com/~std/papers/SIGIR96.ps)

Elliott RJ, L Aggoun, and JB Moore. 1995. *Hidden Markov Models: Estimation and Control*. Springer-Verlag.

Faloutsos C, and DW Oard. 1995. *A Survey of Information Retrieval and Filtering Methods*. Technical Report CS-TR-3514. Dept. of Computer Science, Univ. of Maryland.

Honkela T, S Kaski, K Lagus, and T Kohonen. 1996. "Self-Organizing Maps of Document Collections." *ALMA* (1) 2. (Electronic Journal, address
http://www.diemme.it/~luigi/alma.html.)

Hovy E, K Knight. 1997. "Machine Translation." Course notes from University of California, Department of Engineering, Information Systems and Technical Management Short Course Program. http://www.unex.ucla.edu/shortcourses/spring97/mach.htm

Kipf GK. 1935. *The Psychobiology of Language*. Houghton Mifflin, Boston.

Knight K and V. Hatzivassiloglou. 1995. "Two-Level, Many-Paths Generation." In *Proceedings of the Conference of the Association for Computational Linguistics* (ACL).

Kohonen T. 1995. *Self-Organizing Maps.* Springer-Verlag, Berlin.

Lindgren BW. 1976. *Statistical Theory*. MacMillan Publishing Co, New York.

MacDonald IL, and W. Zucchini. 1997. *Hidden Markov and Other Models for Discrete-valued Time Series*. Chapman & Hall.

MATLAB, Ver. 5.0. The MathWorks, Inc. Natick, MA.

Mosteller F, and DL Wallace. 1984. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*. Springer-Verlag, New York.

Roderick JA. *Statistical Analysis with Missing Data*. 1987. John Wiley & Sons, New York.

Salton G. 1971. *The SMART Retrieval System – Experiments in Automatic Document Processing.* Prentice-Hall Inc, New Jersey.

Samuelsson C. 1996. "Relating Turing's Formula and Zipf's Law." In *Proceedings of the 4th Workshop on Very Large Corpora*, pp 70-78, ACL. Also available as CLAUS Report 78; cmp-lg/9606013.

S-PLUS, Ver 4. Statistical Sciences, Inc. Seattle, WA.

Turtle H, and WB Croft. 1992. "A Comparison of Text Retrieval Models." *The Computer Journal*, 35(3):279-290.

United Nations Parallel Text Corpus (English, French, Spanish). 1994. Available from the Linguistic Data Consortium, http://www.ldc.upenn.edu/ldc/catalog/html/text_html/unptc.html.

Wise JA, JJ Thomas, et al. 1995. Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. *Proceedings of the IEEE '95 Information Visualization*, IEEE Service Center, 51-58. Atlanta GA.

Yarowsky D. 1995. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196. Cambridge, MA.

Personnel Receiving Pay from This Project

KJ Adams	LM Curtis	TJ Martin	DM Rice
DJ Bates	DS Daly	DL McQuerry	IE Roberts
SJ Bohn	SL Eaton	IC McVeety	SK Stade
M Boling	DK Gemeinhart	ES Mendoza	WT Valdez
AJ Calapristi	WM Harris	PA Meyer	PD Whitney
BJ Cheney	EG Hetzler	NE Miller	MV Whyatt
MH Cliff	RW Little	GC Nakamura	J York
KA Cook	CA Lopristi	DA Nielson	CR Younkin
NB Corrigan	EM Lopristi	KA Pennock	
VL Crow	AE Lundeby	MC Pottier	

APPENDIX A

TRAUMA DATA ANALYSIS

SPIRE

Trauma Data Analysis

Project Manager: Elena Mendoza
Group Manager: Renie McVeety
Department Manager: GL Work

Prepared for the

U.S. Army Medical Research and Materiel Command
with the U.S. Department of Energy
Contract DE-AC-06RLO 1830

October, 1996

Table of Contents

Introduction

The US Army MRMC is interested in applying Spatial Paradigm for Information Retrieval and Exploration (SPIRE) technology to data sets of trauma related information. SPIRE has generally been designed to work with larger data sets of a more unstructured nature (i.e. newswatch data, message trafficking). The purpose of this research was to determine if SPIRE would produce meaningful analysis of this trauma data. The trauma data provided by Dr. Howard Champion of the University of Maryland, has been processed and analyzed. The data consisted of 86 documents from two different sources pertaining to auto accidents. Both sources of documents discuss traumas; however, one set discussed more of the accidents, while the other source discussed more of the medical problems. SPIRE identified the two major differences in data sources, but, according to our analysis, could find little other meaningful differentiation among the documents. From our initial analysis, it appears as though the low number of documents and the structured nature of those documents impacted the performance of SPIRE on this data set.

Task Description and Study Results

SPIRE technology, as it exists today, assesses the patterns of word use in documents (e.g. frequency clusters of an individual word implies salience in theme) to determine important topics and relationships. Natural language communication utilizes definite strategies for conveying the content particularly when substantial knowledge is not assumed. SPIRE is dependent on this. Because of the inherent structure in the trauma data, the SPIRE system has the tendency to identify most documents in this data set as very similar. As a consequence, thematic differentiation of this data set is less pronounced than we would normally consider desirable.

Several different analyses were generated to assess the nature of this document set. First, we examined the correlation matrix (which establishes the correlations between all major terms and the key topics), and found consistently lower than normally acceptable values. A subset of the matrix can be found in **Appendix A: Figure 1**. A random sample of terms is provided to convey the nature of the correlation matrix content. The term at the top of each column is the term to which other terms are related. The values are normalized and in a more strongly related set of documents may average 0.6 or higher. As the sample conveys, the average for the trauma data is about 0.15 for the ten most related terms. Additionally, the number of connected terms in a "normal" data set is typically much greater (terms with non-zero values), which again demonstrates the structured nature of this data. Generally, the information found in this matrix indicates that the documents don't tend to group into easily differentiated clusters and that relationships which would otherwise be small enough to ignore, can have a dominant influence on thematic distribution.

A second analysis we performed was to identify the number of statistically important topic terms in the data set relative to the unique word count. This is commonly called the noise to signal ratio, the "signal" being the important topics and the rest of the vocabulary being the "noise". There were only 13 major topics found in the data set with 2806 words in the vocabulary. This percentage (.005%) is extremely low for significant analysis. Significantly fewer high-value topical terms were found in this data set than were found in more normal expressions of natural language communication such as news articles or research papers or even WWW pages. For example, a data set run on CNN news data (635 small documents) gave approximately 450 major terms and 10,000 words in the vocabulary for a percentage of 4%. The trauma data set possesses a high noise to signal ratio, making it difficult to find meaningful information.

Appendix A: Figure 2 is another illustration of the lack of discriminating dimensions in the data set. In this graph, there are three documents which appear in very different locations in the 2D scatter plot: two are proximal, the third is distant from the first two which is shown by **Appendix A: Figure 3**. Thirteen topics along with their relative magnitudes are graphed. Typically, we would expect to be able to quickly identify proximal and distant document pairs due to the strong diversity in content represented by the magnitudes of each dimension--the relative magnitudes at each dimension of proximal documents would be close while the values for distant documents would be quite different. A couple of dimensions or topics in this data set, "pilon" and "travel" follow this pattern. However, the other topics show more random values for both proximal and distant pairs. This again shows that the data is not rich in discernible content, at least as measured by SPIRE.

There were, however, some positive results that are worthy of note. We were able to identify thematic clusters which could provide some insights to an analyst, depending upon their domain-specific requirements. For example, a query on the word, "tibia," shows that all tibial fractures tend to group together in the lower middle quadrant of the themescape as shown in **Appendix A: Figures 4 and 5**. Further exploration with a larger data set might enable us to discover correlations with such elements as vehicle type or speed, age of "case," and so on.

Conclusions

In summary, what we've determined is that this particular data set has a degree of structure which makes it difficult for the SPIRE system to meaningfully process. Each document discusses the same general topics in the same general language; therefore, the language used doesn't convey the importance of words in the manner to which SPIRE is tuned. There is, however, information in the structure itself that has potential. Our domain expert, Dr. Howard Champion, agrees with this analysis and believes that SPIRE accurately portrayed the information in the trauma data.

At least two short term steps might be taken to improve the quality of the results. First, acquiring a larger data set might yield a better correlation matrix in terms of strength and number of word/word correlations, although we are skeptical that more data of exactly the same type would produce appreciably better results. Second, it might be possible to separate or eliminate some of the non-injury related vocabulary in the belief that the remaining data will map out more meaningfully. This effort would eliminate some of the "noise" in this signal.

Further research into this type of data and how to process and visualize it in a meaningful way is required for more substantive progress. The field of visual analysis of structured data is new and innovative. Research would include understanding the structure and gaining knowledge from the structure. It would also include new ways of visualizing structured data that might combine current data mining techniques with new visualization techniques such as SPIRE. SPIRE could be expanded to support this area in conjunction with it's current architecture for dealing with unstructured text.

Appendix A: Figures and Charts

Figure 1. Correlation Matrix

The following matrices show a random sample of 12 terms and the normalized correlation value of 10 related terms. Strongly related document sets average 0.6. Trauma data average is approximately 0.15.

Term: 1st

Related Terms:		Normalized correlation value:
	victim	0.266667
	embankment	0.266667
	plateau	0.133333
	pilon	0.133333
	concrete	0.133333
	tibial	0.0666667
	pole	0.0666667
	metatarsal	0.0666667
	airbag	0
	travel	0

Term: 5th

Related Terms:		Normalized correlation value:
	metatarsal	0.266667
	cuboid	0.2
	embankment	0.2
	bimalleolar	0.133333
	airbag	0
	lisfranc	0
	victim	0
	travel	0
	tibial	0
	pilon	0

Term: 3rd

Related Terms:		Normalized correlation value:
	lisfranc	0.266667
	cuboid	0.266667
	victim	0.266667
	metatarsal	0.2
	airbag	0.0666667
	travel	0
	bimalleolar	0
	embankment	0
	pole	0
	pilon	0

Term: accelerator

Related Terms:		Normalized correlation value:
	travel	0.0666667
	tibial	0
	plateau	0
	pole	0
	concrete	0
	pilon	0
	victim	0
	metatarsal	0
	cuboid	0
	embankment	0

Term: 4th

Related Terms:		Normalized correlation value:
	metatarsal	0.333333
	cuboid	0.2
	embankment	0.2
	pilon	0
	pole	0
	concrete	0
	tibial	0
	plateau	0
	bimalleolar	0
	lisfranc	0

Term: access

Related Terms:		Normalized correlation value:
	bimalleolar	0.133333
	cuboid	0
	metatarsal	0
	embankment	0
	airbag	0
	lisfranc	0
	victim	0
	travel	0
	tibial	0
	plateau	0

Term: accident

Related Terms:		Normalized correlation value:
	airbag	0.0666667
	metatarsal	0.0666667
	plateau	0
	tibial	0
	pole	0
	concrete	0
	pilon	0
	victim	0
	bimalleolar	0
	cuboid	0

Term: airborne

Related Terms:		Normalized correlation value:
	victim	0.266667
	embankment	0.266667
	concrete	0.266667
	metatarsal	0.2
	plateau	0.133333
	airbag	0.0666667
	tibial	0.0666667
	cuboid	0
	bimalleolar	0
	pilon	0

Term: acetabulum

Related Terms:		Normalized correlation value:
	cuboid	0.133333
	victim	0.133333
	pilon	0.133333
	metatarsal	0.133333
	plateau	0.133333
	pole	0.0666667
	tibial	0.0666667
	travel	0
	lisfranc	0
	embankment	0

Term: apparently

Related Terms:		Normalized correlation value:
	embankment	0.4
	cuboid	0.4
	pilon	0.2
	metatarsal	0.133333
	tibial	0.0666667
	concrete	0.0666667
	victim	0.0666667
	lisfranc	0.0666667
	pole	0
	airbag	0

Term: airbag

Related Terms:		Normalized correlation value:
	airbag	1
	victim	0.2
	cuboid	0.2
	metatarsal	0.133333
	tibial	0.0666667
	pole	0
	pilon	0
	plateau	0
	travel	0
	lisfranc	0

Term: arm

Related Terms:		Normalized correlation value:
	airbag	0.2
	cuboid	0.2
	victim	0.2
	tibial	0.0666667
	bimalleolar	0.0666667
	pole	0.0666667
	travel	0
	plateau	0
	lisfranc	0
	concrete	0

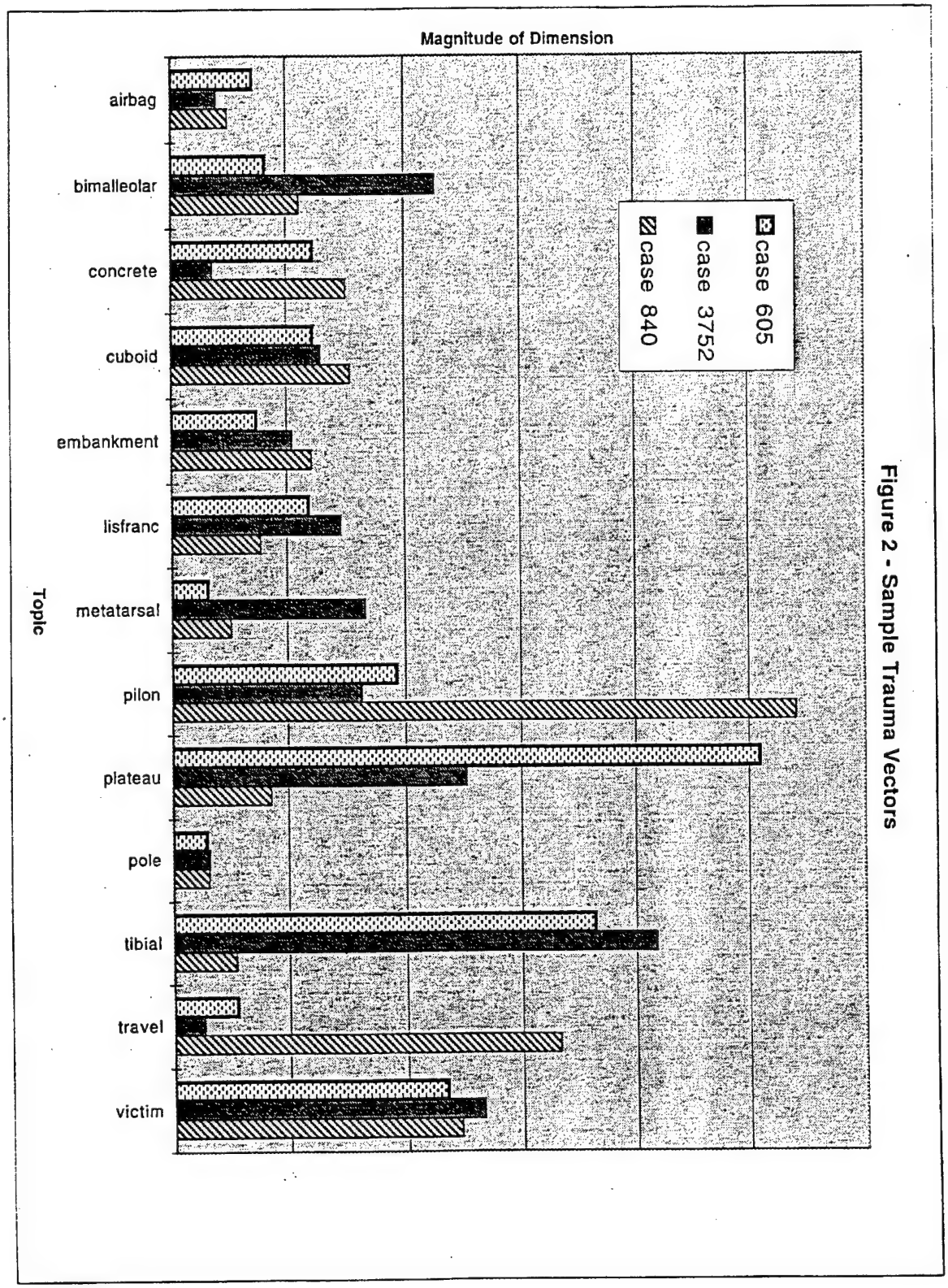
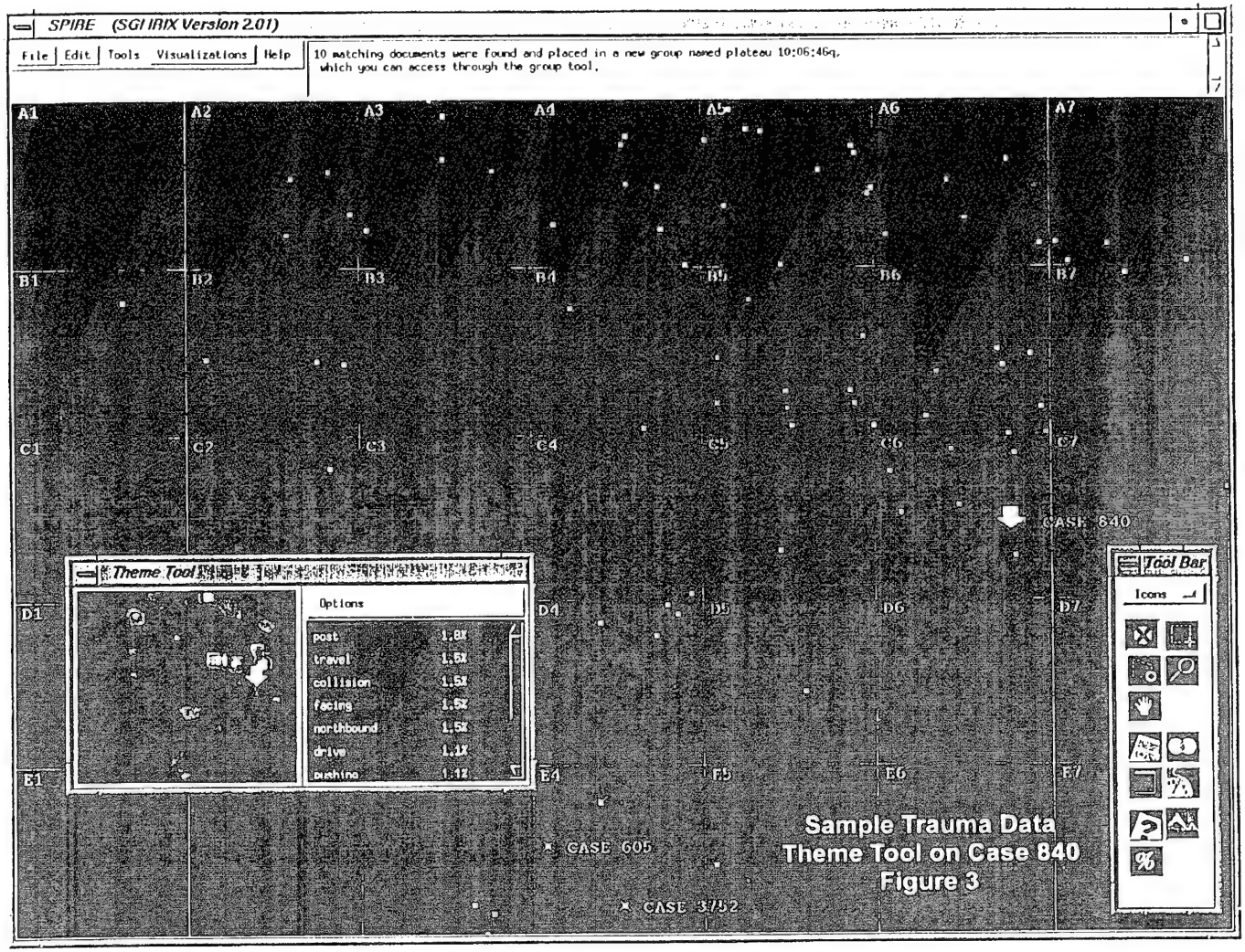
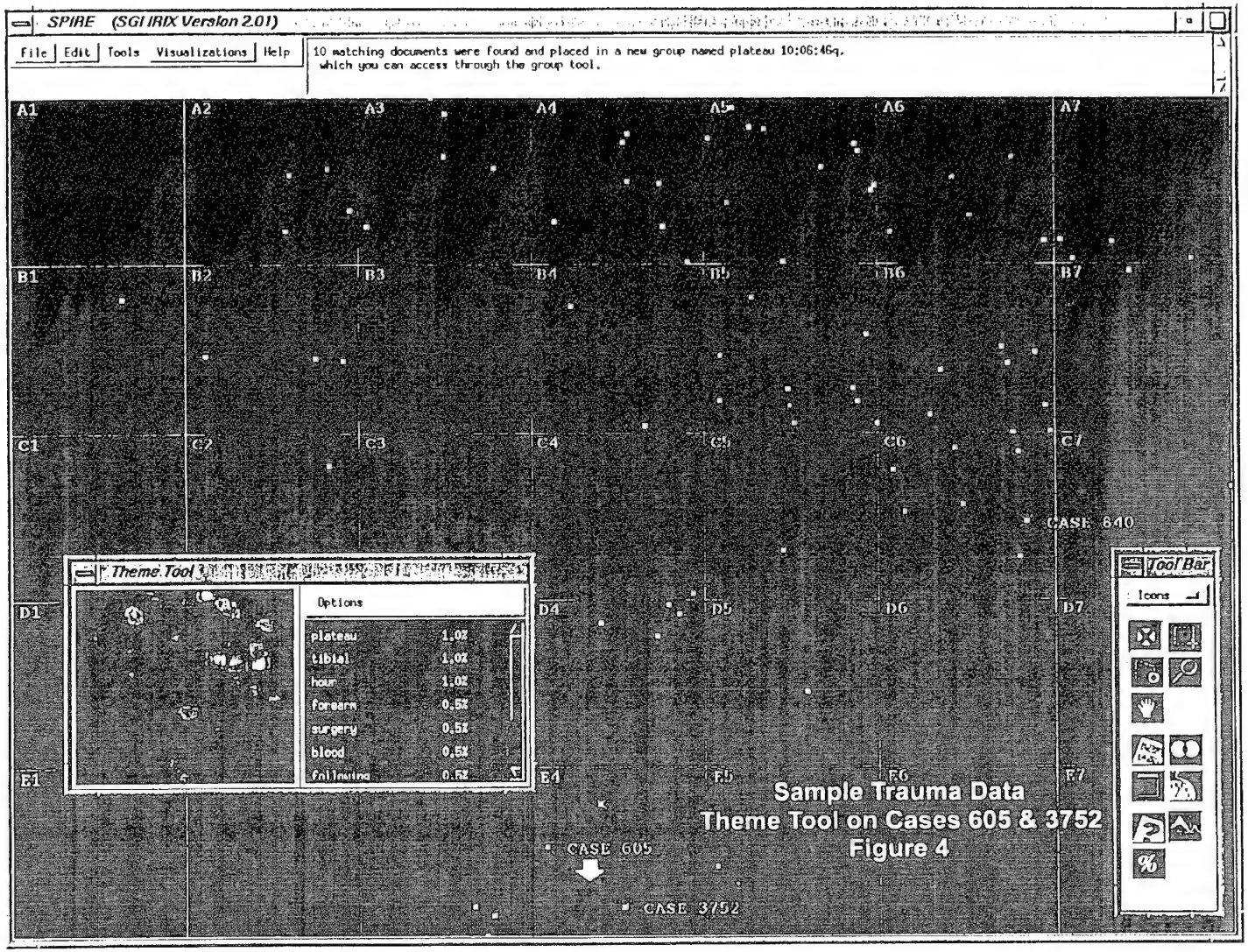


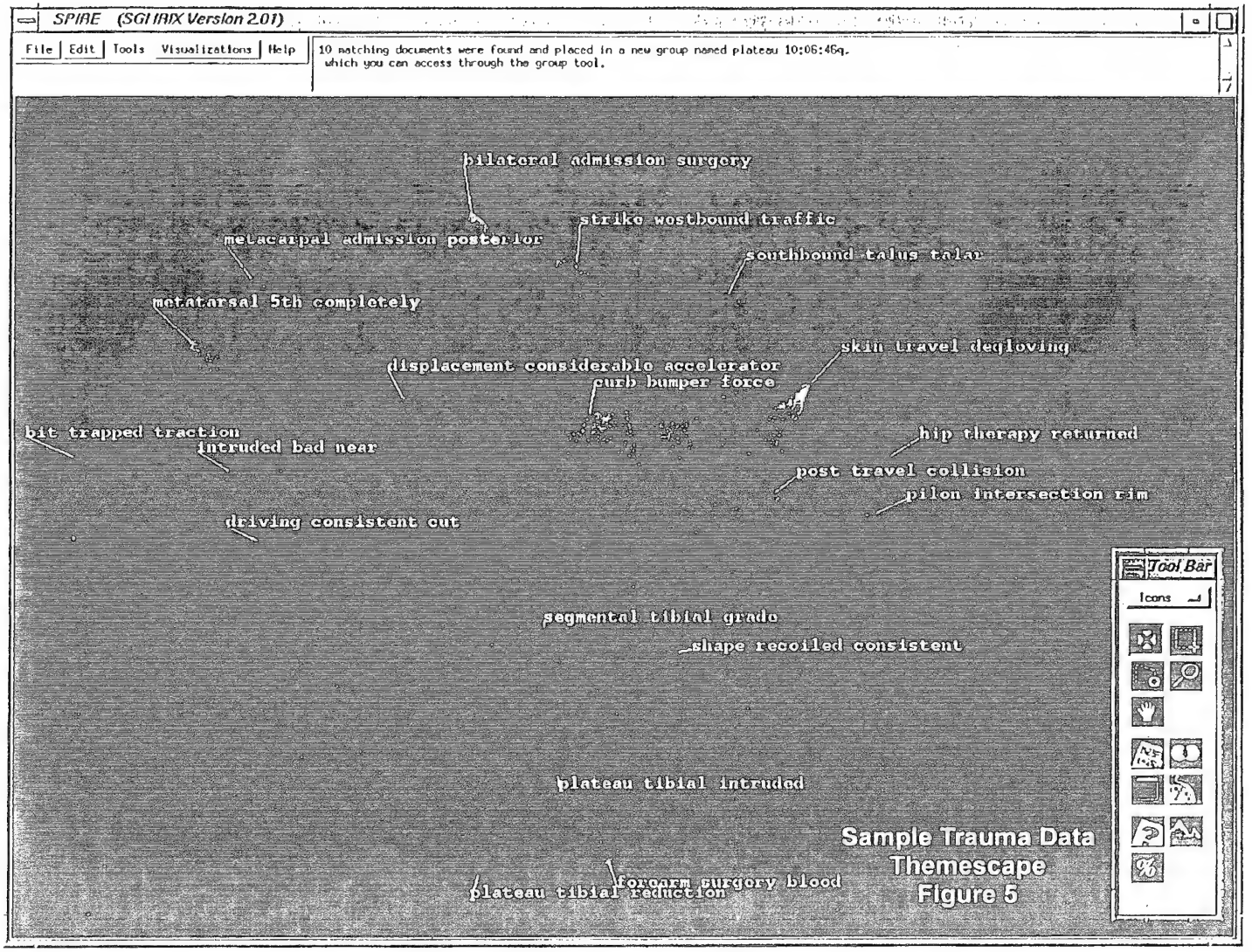
Figure 2 - Sample Trauma Vectors



Sample Trauma Data
Theme Tool on Case 840
Figure 3



Sample Trauma Data
Theme Tool on Cases 605 & 3752
Figure 4



Sample Trauma Data
Themescape
Figure 5

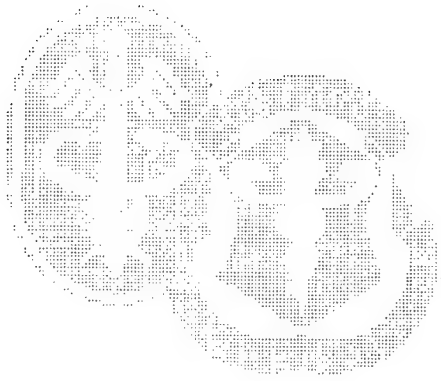
APPENDIX B

VISUALIZATION RESEARCH AND DEVELOPMENT:

STRUCTURED AND UNSTRUCTURED TEXT ANALYSIS

Information Analysis
Visualization Research and Development

Structured and Unstructured Text Analysis



Project Manager: A.J. Calapristi
Group Manager: Renie McVeety
Department Manager: GL Work

Prepared for the

U.S. Army Medical Research and Materiel Command
with the U.S. Department of Energy
Contract DE-AC-06RLO 1830

July 2, 1997

Structured and Unstructured Text Analysis

Table of Contents

1.0 Introduction

The U.S. Army Medical Research and Material Command (MRMC) has commissioned Battelle to conduct research in applying SPIRE (Spatial Paradigm for Information Retrieval and Exploration) to structured textual information. This report addresses Task 1, "Processing and Analysis of Planning Documents", and Task 2 "Processing and Analysis of Structured Data sets" in the Information Technology Proposal. Specific areas of research included

- Determine if SPIRE's advanced statistical analysis and graphical presentation techniques are valid for processing structured text
- Evaluate approaches to improve the processing and presentation of structured text.
- Research approaches to correlating and relating different documents sets

2.0 Challenges

2.1 Structured Text Context

SPIRE technology was initially developed for the intelligence community and was designed to work with large sets of unstructured text files such as message traffic, news wire feeds, etc. While structured text may be written in full sentences like a news wire feed, the actual meaning of the text relies more heavily on context than unstructured text. Capturing the context of the structured text is a major challenge in processing the MRMC data.

2.2 Large Documents-Small Quantity

The MRMC provided thirty-four data files containing an assortment of scientific/technical planning and management information. These included Accomplishments by Program element, Scientific and Technical Objectives, and Thrust Areas. The total number of documents received from the MRMC was too small to provide a statistically valid sample set for SPIRE processing. Additionally, many of the document files received for processing contained complicated structuring, and/or summary level text. The formatting style of most of the documents was similar to those used in view foil presentations.

2.3 Visual Comparison of Document Sets

SPIRE is typically used to analyze collections of documents from a single source such as a newswire feed or a specific information service. Applying this technology to comparison of documents from different sources (e.g., planning documents from different organizations) seems to be a natural extension. Although SPIRE can organize the documents based upon their thematic content, it does not offer a function that enables the user to discriminate which documents came from which source. This capability is essential enabling the user to make visual comparisons.

2.3 Project Scope

This effort covers Tasks 1 and 2 in the Statement of work. The research and understanding of applying SPIRE to structured text processing will be significantly advanced though the MRMC effort. However, the design and engineering of the software to implement the research findings will be significant, and is outside the scope of these tasks. A major challenge for the project team is to keep the priority on research and technology advancement, and where possible minimize costs associated with software modifications.

3.0 Approach

A series of workshops were conducted to analyze the technical issues surrounding structured and unstructured data analysis, and to determine the road map for addressing the MRMC task objectives. These sessions resulted in the overall process outlined below, and established the initial technical directions that were pursued.

1.0 Overall Approach

- Research generalized approaches for structured text analysis and comparing sets of documents (i.e., planning vs. execution) within a Galaxies projection. Use data sets currently available at PNNL for the research.
- When received, analyze the MRMC data sets for structure, content, and applicability to technical directions resulting from research and brainstorming.
- Process the MRMC raw document sets to produce baseline Galaxies and Themescape visualizations.
- Analyze the baseline document sets to assess the validity of the document clustering.
- Where feasible, apply these new approaches and generate new visualizations of the data for comparison with the baseline.

3.2 Initial Technical Directions

The following approaches were considered for addressing the structured data and document comparison tasks.

3.2.1 Document Comparison using a multi pass analysis.

This method would have used a first pass analysis to establish the thematic content and term associations expected in subsequent document sets, and subsequent analysis passes to quantify the differences among the documents.

For example, in the first pass the planning document corpus would be processed to define the expected relationship among the major topics. The secondary passes would processes the supporting work execution documents (e.g., structured text), using the relationships established in the first pass, as the basis for the Galaxies or Themescape display. Each of the secondary passes would process a statistical best fit to the document set used in the first pass. Essentially, this approach uses the first pass to "learn" what the associations

should be, and the second pass to map the work execution documents based upon the planning document analysis. The degrees of Galaxies clustering and/or Themescape elevation will indicate how well the two document corpora matched in topicality.

3.2.2 Document Segmentation for Structured Documents

This approach involves re-developing the current text analysis approach to divide individual documents into many smaller documents segments (i.e., virtual documents). This could be done by user-defined break points (e.g. paragraphs or specific sections) or through some software mechanism that could automatically identify thematic breaks in the text. Essentially this would create a document corpus of virtual documents derived from the real documents. The Galaxies or Themescape projection would map each virtual document on the display but enable the user to link back to the real document source.

The advantages to this approach include:

1) document corpora of different types could be processed together since the real documents would be standardized during the segmentation into virtual documents;
2) the original document could be located on the display in as many places as it has document segments.
3) better analysis of corpora which contain long documents or documents which cover a broad and complex range of themes.
4) ability to analyze a single large document or small number of large documents.
5) facilitation of comparisons (e.g. plans vs. actions).

Three methods of document segmentation were discussed:

Statistical Context Analysis uses wavelet analysis to "sense" where thematic breaks occur in a document, and segments the documents on these break points.

Discrete Segmentation divides documents up based upon discrete breakpoints such as number of characters, words, punctuation, or breakpoints.

User Defined Context method applies user-defined phrases or keywords to identify where documents segments should begin or end.

3.3 Selected Technical Approach

The technical approach is summarized below.

3.3.1 Document Segmentation

The User Defined Context method was selected to break apart large documents. This addressed three of the major challenges in this project by

1. having the highest chance of retaining the context by ensuring the section headings were tied to the corresponding text,
2. producing a sufficient amount of documents for statistical analysis,
3. and could be accomplished without software modifications to SPIRE.

Statistical Context Analysis, and Discrete Segmentation methods were considered but were not pursued because the state of the research in these areas had not yet advanced to the point it could be easily applied.

The multi-pass analysis method that was discussed in section 3.2.1 was also considered, but it would have required significant changes to the text analysis software. Also, it would have required additional and research and development to create a SPIRE software module that performed the statistical best fit.

3.3.2 Processing and Visualizing Documents from Different sets

The team elected to use a file tagging technique to uniquely identify where the document subsets originated. This could be accomplished when the document subsets were created as described in 3.1.1. The tags could be queried using an existing function in SPIRE that creates document groups, which appear on the screen in different colors. All of the document segments would be analyzed by SPIRE and clustered as though they came from a single source. If there is overlap in the key themes of the planning and execution documents they can be expected to cluster together. If there is not a high degree of commonality in the thematic content, then the documents will most likely cluster according to their source.

3.3.3 New Visualizations

New visualization techniques would be tried using graphics and statistical packages outside the SPIRE environment.

3.4 Analysis of MRMC Document Sets

3.4.1 MRMC Document Taxonomy

The MRMC provided thirty-four data files consisting of planning and management information. The general outline of the files included "Accomplishments by Program Element", "Scientific and Technical Objectives", Thrust Areas, etc.
Twenty of these files were pre-processed to produce 800 "sub-documents". The remaining fourteen files consisted of simple lists, which did not provide enough topical content or unique structure to be segmented into smaller files. In addition, seven Defense Technology Area Plans (TAPS) from Chapter VI, Biomedical were harvested from the Internet and segmented into 20 sub-documents. These were added to the MRMC planning documents database.

Attachment 1 lists the original data sources that were provided by the MRMC, and shows which files were used in the analysis by a * in the first column. Attachment 2 contains an example of a SPIRE source file containing sub-documents.

3.4.2 Validation Methods for the Visualizations

Two methods were used to determine the if the visualizations produced meaningful results. Part of the analysis effort involves "tuning" parameters such as topicality indexes to get closer to target ranges, and adjusting other parameters to account for low word counts, highly technical vocabularies, or small document sets.

Topicality Ratios: Based upon past analysis efforts, a reference point of 100:10:1 has been established for the ratio between word:count:cross terms:topics. A starting topicality index of 2 and a cross term index of 1 are used as the initial values to determine topics and cross terms. Document corpora that fall close to the above parameters usually produce visualizations that are well differentiated and meaningful to the analyst. However, if the topicality index has to be moved too low (below 2), the differentiation of documents in the visualization may be compromised.

Qualitative Evaluation: The second method, and often the most reliable, is a subjective measure based upon analysis of the data in the projection.

Qualities to consider are:

1. Document Spread- are their sections of the visualization where the documents are too tightly clustered for the user to effectively analyze them?
2. Topical cohesion – Do the various regions of the display seem to be sensibly organized or are some topics too dispersed?
3. Use of Space – are their broad areas of the visualization that contain no data, and does this "dead space" convey any useful information?

Based upon the above criteria, a subjective measure of Very Good, Good, Fair, or Marginal is assigned to each of the MRMC projections.

3.4.3 Visualizations

3.4.3.2 Baseline Visualization Evaluation

Attachment 3 shows a visualizations of the 20 MRMC files before they were segmented into document subsets. The following table is a summary of the evaluation metrics for The Army Baseline document corpus.

Army Baseline – No Segmentation

Topic Index: 3.6
Cross Term Index: 2.3
Documents: 20

Measure	Value
Topicality Ratio	100>3.3>2.3
Qualitative	Marginal

The Topicality Ratio indicates many unique or specialized terms were used in the corpus. This was a principally a result of the high use of bulleted phrases, abbreviations, and short summary paragraphs used in the documents. This type of high ratio would be expected since the authors of the documents were attempting to convey the most information with

the least amount of words. In addition, the documents used precise language in very short descriptions of programs from many different scientific and medical disciplines. This phenomenon is not uncommon when dealing with sparsely worded technical documents like those found in the MRMC corpus.

Although topical cohesion of the documents was high, The qualitative evaluation of the visualization is marginal because the low sample set did not provide enough documents for a helpful visualization of the document content. Much of the topical content remained hidden within the 20 documents.

3.4.3.3 Document Segment Visualizations

Attachment 4 shows Galaxies and Themescape visualizations of the ASBREM and RAD corpora. For the most part, the document segments derived from these data sets produced a valid statistical sample size for the visualizations. However, the original data files were not always consistent in their structure which resulted in some documents containing very sparse or duplicate textual descriptions. These anomalies did not appear to significantly affect the visualizations. The following tables are summaries of the evaluation metrics for these visualizations.

Note: "Documents" indicates the total number of document segments that were produced from the original source file.

RAD I-VI

Topic Index: 2
Cross Term Index: 1
Documents: 554

Measure	Value
Topicality Ratio	100>14>2
Qualitative	Good

ASBREM 93

Topic Index: 2
Cross Term Index: 1
Documents: 87

Measure	Value
Topicality Ratio	100>15>3
Qualitative	Good

ASBREM 95

Topic Index:	3.5
Cross Term Index:	1.5
Documents:	159

Measure	Value
Topicality Ratio	100>21>3
Qualitative	Fair

The document segmentation produced document corpuses with topicality ratios closer to the reference point of 100:10:1. This was a principally a result producing smaller documents that addressed a much more specialized focus. Although the terms used within the entire document corpus did not change, the text analysis software was able to create a greater number of document signal vectors containing highly specialized numerical descriptions for each document segment. The ASBREM 95 documents used a vocabulary that contained a very high ratio of statistically important terms to total word count. Even when the topic index and cross term indexes were increased the topicality ratio remained significantly higher than the reference point.

The qualitative evaluation of the visualizations ranged from Good to Fair. While this is an improvement over the Army Baseline corpus, the low numbers of documents in each corpus affect the quality of the visualizations.

3.4.3.3 MRMC Planning Documents

Attachment 5 shows Galaxies and Themescape visualizations of the combined 800 document segments files produced from the MRMC planning documents, and 21 segments produced from the Army Biomedical Technical Area plans. One of the visualizations shows groupings by program area. The color legend is in the upper right hand portion of the visualization (note: green dots designate TAP documents). The following table is a summary of the evaluation metrics for the MRMC Planning Documents.

MRMC Planning - all

Topic Index:	2
Cross Term Index:	1
Documents:	820

Measure	Value
Topicality Ratio	100>14>1
Qualitative	Very Good

This document corpus contained enough documents and topical diversity to provide a very good topicality ratio.

The qualitative evaluation of the visualizations was very good and provided useful information regarding the relationships of documents from various programs and scientific areas. An interesting paradox is that the rich vocabulary helped to differentiate the information in these sparsely worded documents, while it tended to have a negative effect when the information existed as a single document. As with the smaller document sets, there appeared to be five major thematic areas that emerged in the visualizations. These included:

1) Information related to Infectious Disease, which included information on viruses, vaccines, and parasitic diseases such as malaria, drugs and antigens.
2) Chemical and Biological Defense including information on toxins, agents, and countermeasures.
3) Information Related to Human Performance, including environmental stress factors
4) Injury, trauma care and other conditions related to the battlefield
5) Dental programs including, prevention and treatment of dental disease, maxillofacial injury, and diagnostic tools such as new x-ray machines.

The visualization also showed the relationship among the Biomedical Technical Area Plans (TAPs) and the MRMC information. There was excellent correlation between the TAPS and the MRMC plans. For example all thee areas from the BioMedical Combat Casualty Care TAP was located in the same cluster that contained JTCG6, and RAD ii, data related to injury treatment.

3.4.3.4 Time Slicing

Attachment 6 contains visualizations that show subsets of the MRMC Planning documents in 1993, 1994 and 1995. The bar chart on the time slicer shows the corresponding time segment for the visualization highlighted in blue. The bar chart also shows the relative number of documents for each year, and their corresponding group.

Generally, the results of the time slice analysis showed an even distribution of information among all four major thematic areas in the 1993, with the notable exception of the chemical warfare information. The 1994 data contained the most documents and showed a significant increase in data associated with chemical and biological warfare, and infectious disease. 1995 data showed relatively more information in the areas of dental and human performance categories.

4.0 Conclusion

Structured text, like that represented by the MRMC can be processed into useful visualizations. The consolidated MRMC Planning document set provides an example of how SPIRE technology can be used to quickly analyze large numbers of documents, and provide new insights regarding relationships among information.

Once a subject matter expert becomes familiar with the analytical capabilities of SPIRE, they can sort through the information before deciding which documents are most relevant and discover the content within large text document sets with minimal to no reading of the

documents. This approach can be applied to areas of medical research, scanning of data from government information sources such as DTIC, and for analysis of data sets from a researchers personal electronic file.

The true value of SPIRE is realized when a user can interact with the system to provide fresh insights into their data that may not have been apparent using traditional text analysis approaches. SPIRE has been installed at WRAIR. A good first step would be to install SPIRE on a computer at Fort Detrict. Preliminary contacts have been made with the Directorate of Information Management (DOIM) to identify a suitable computer system.

The second step is to continue to advance the software technology of the SPIRE application so that it can be accessed from a greater number of users. In the case of the MRMC this requires changes to enable the SPIRE software to run on PC computers. The preliminary work to establish a conceptual architecture for the next generation of SPIRE has been completed. Additional funding is required to refine the design, and proceed with the development of the new software.

A third area is the addition of new functionality such as data harvesting, automated segmentation of large documents, the ability for users to pre define topics for visualization, support for structured fields that can be used to create additional subsets. Work has already been started in these areas, but needs additional sponsorship to continue.

Attachment 1

MRMC Data Files

	DISK	File	tag	Description
	'93 ASBREM S&T's			**Armed Services Biomedical Research Evaluation and Management (ASBREM) Committee**
*		Mildent93	mildent	Military Dentistry
*		Infdis93	Infdis	Infectious Disease
*		humsysa93	Humsys	Human Systems Technology
*		humsys93	Humsysa	Human Systems Technology
*		chemdef93	Chemdef	Medical CW Agent Defense
*		Ccc93	Ccc	Combat Casualty Care
*		Biodef93	Biodef	Biological Warfare Defense
	95 ASBREM S&Ts			**Armed Services Biomedical Research Evaluation and Management (ASBREM) Committee**
*		Jtcg1-95	Mildent	Military Dentistry
*		Jtcg2-95	Parasitic	Infectious Disease
*		Jtcg4-95	Biodef	Biological Threat Agent
*		Jtcg5-95	Opermed	Human Performance (Stress)
*		Jtcg6-95	Ccc	Mission (Far Forward Care)
*		Jtcg7-95	Nuclear	Nuclear Weapons
	R&A 94 TAP			**Defense Technology Area Plan for Biomedical Science and Technology**
*		Tap.txt		TAP overview and table of contents
*		Radi.txt	Radi	Infectious Diseases of Military Importance
*		Radii.	Radii	Combat Casualty Care
*		Radiii	Radiii	Army Systems Hazards
*		Radiv	Radiv	Medical Biological Defense
*		Radv	Radv	Medical Chemical Defense
*		Radvi	Radvi	Breast Cancer Research
	Tara96			
		Tarazimn		Military Telemedicine
		Tarabrf		Telemedicine Testbed
		Radiolog		Radiation
		Mom		Focus on the WARFIGHTER
		Infectio		Infectious Disease
		Dental		Dental
		Chemical		Chemical
		Ccc		Combat Casualty Care
	Battle Labs 96			**Research area title plus STO cross reference in Paraenthesies**
		Blintro		Introductory Material
		Blst1		Infectious Disease
		Blst2		Far Forward Care
		Blst3		Battle Effects
		Blst4		Biological Countermeasures
		Blst5		Biological Countermeasures

Attachment 2

Example Source File

The data following the dotted line is an excerpt from the source file used to produce the MRMCPlan visualizations. The format of the file is shown below .

Record Delimiter: **RECORDKEY**
Title Identifier: **TITLE:**
Date Identifer **DATE:**
Tag: **MRMC radii (changes for each type of source document such as radi, radiii, etc)**

```
RECORDKEY
TITLE: [radii] HIGHLIGHTS: Established model for uncontrolled extremity
hemorrhage and used model to
DATE: 1994
Established model for uncontrolled extremity hemorrhage and used model
to
conduct proof-of-concept testing for fibrin glues as local hemostatic
agents.
IMPACT:     Fibrin glues have the potential to be used as a soldier-
level item
to improve far-forward
control of extremity hemorrhage, a significant preventable cause of
death on the
battlefield.
OPPORTUNITIES:    Additional studies will be performed to optimize
fibrin glue
formulations and
obtain further evidence of efficacy and safety.  The model of
uncontrolled
extremity hemorrhage will be
used as a reference model for other studies of resuscitation and shock
intervention.
MRMC radii

RECORDKEY
TITLE: [radii] HIGHLIGHTS: Demonstrated that silver nylon dressings
promote
healing of donor site wounds
DATE: 1994
Demonstrated that silver nylon dressings promote healing of donor site
wounds
and provide patient comfort.
IMPACT:     Local sepsis is a major complication of burn wounds and is
also a
factor that can lead to
incapacitation and a requirement for evacuation of soldiers after
relatively
minor mechanical trauma.
Improved wound dressings have potential to reduce hospital costs,
improve
patientsÕ quality of life, and
keep soldiers on the battlefield.
```

OPPORTUNITIES: Silver nylon dressings for burn care will be transitioned to development, and the potential of silver nylon for treatment of ballistics wounds will be explored.
MRMC radii

RECORDKEY
TITLE: [radii] HIGHLIGHTS: Developed three-week suspended release formulation for microencapsulated
DATE: 1994
Developed three-week suspended release formulation for microencapsulated cephazolin.
IMPACT: Up to 70 percent of all soldiers wounded in action sustain wounds to the extremities, mainly due to ballistics. Development of biodegradable microencapsulated antibiotics will provide improved control of debilitating osteomyelitis (bone infection) resulting from such wounds, resulting in improved quality of life for injured soldiers and reduced care costs.
OPPORTUNITIES: Further studies will be performed to develop formulations that provide four to six week release of cephazolin and tobramycin antibiotics and demonstrate their efficacy and safety in animal models. These products will be transitioned to development as a complement to microencapsulated ampicillin, which is already in development, thereby providing broad spectrum antimicrobial activity.
MRMC radii

RECORDKEY
TITLE: [radii] HIGHLIGHTS: Continued production of Òpilot scaleÓ batches of stroma-free hemoglobin.
DATE: 1994
Continued production of Òpilot scaleÓ batches of stroma-free hemoglobin.
IMPACT: Availability of a blood substitute will allow oxygen-carrying resuscitation fluids to be administered at Echelons 1 and 2, where logistics considerations largely preclude the use of fresh or freeze-thawed refrigerated blood products. Army hemoglobin production permits investigations of the efficacy and toxicity of candidate blood substitutes by independent academic researchers, unhindered by restrictions on information exchange that have been imposed by commercial blood substitute developers.
Such investigations are deemed crucial by the National Institutes of Health (NIH) and the FDA to ensure that issues of safety and efficacy with these products are comprehensively addressed in a timely manner.

OPPORTUNITIES: Stroma-free hemoglobin will continue to be produced and
provided to qualified
academic investigators proposing research of high programmatic relevance, and
will be provided to U.S.
Navy investigators to facilitate their efforts in developing liposome-encapsulated hemoglobin as a
alternative blood substitute material.
MRMC radii

RECORDKEY
TITLE: [radii] HIGHLIGHTS: Designed a suite of life-support medical equipment
for battlefield resuscitation
DATE: 1994
Designed a suite of life-support medical equipment for battlefield resuscitation
of trauma victims, resulting in three patent disclosures
IMPACT: Development of a closed loop autonomous life-support system for far-
forward
resuscitation will allow medics and forward surgeons to provide an improved
standard of care on the
battlefield and during evacuation, and will reduce the time that medical personnel need to attend to each
casualty, enhancing capabilities for mass casualty management.
OPPORTUNITIES: Collaborative efforts with academia and industry will continue
to develop
improved diagnostics, adapt existing device technologies, and integrate these
into prototype systems

Personnel

FY93 END STRENGTH
SUMMARY

FY93 END STRENGTH BY CATEGORY

RAD II Funding Summary

FY93 Funding

MRMC radii

Attachment 3

Visualizations

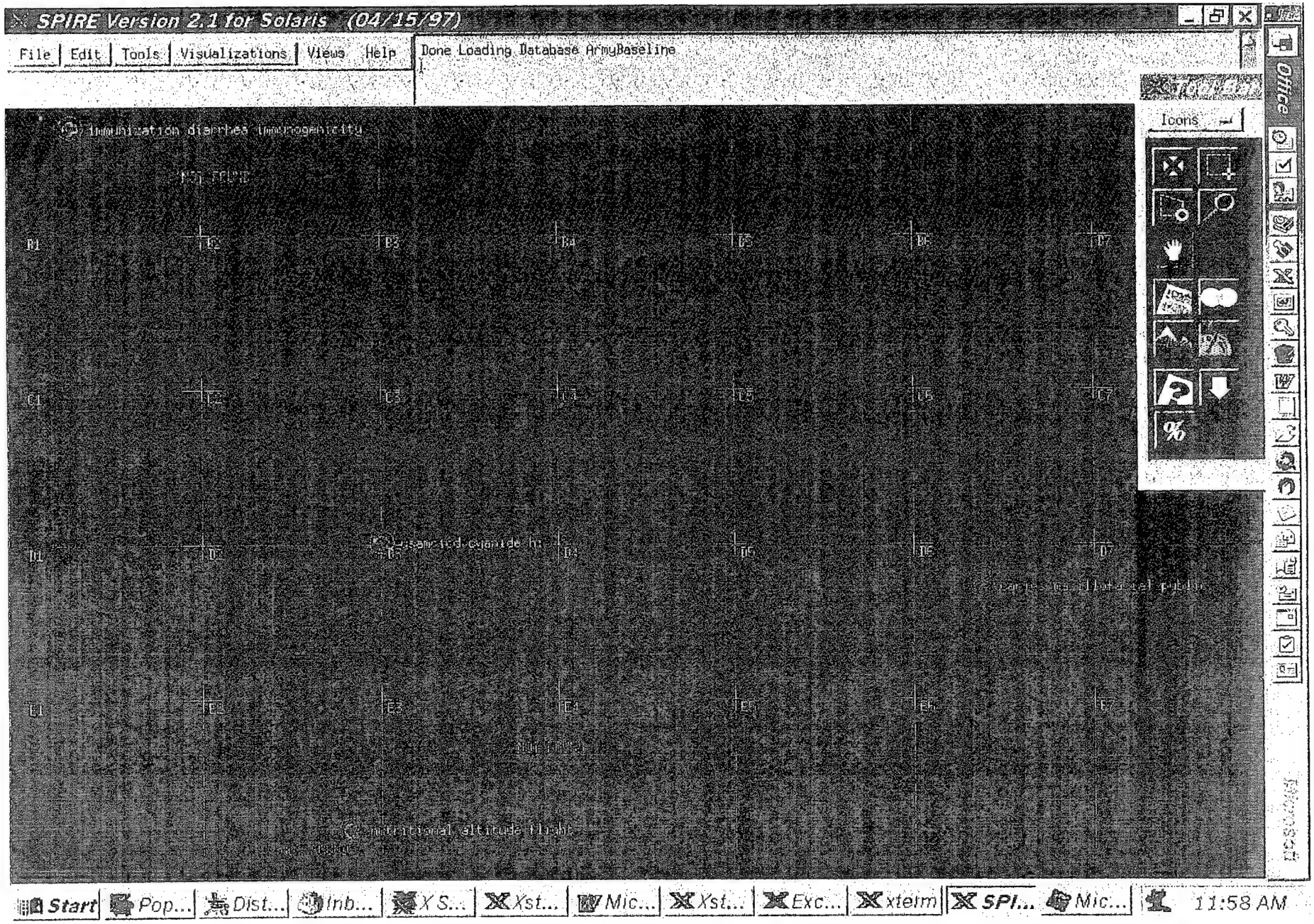
of

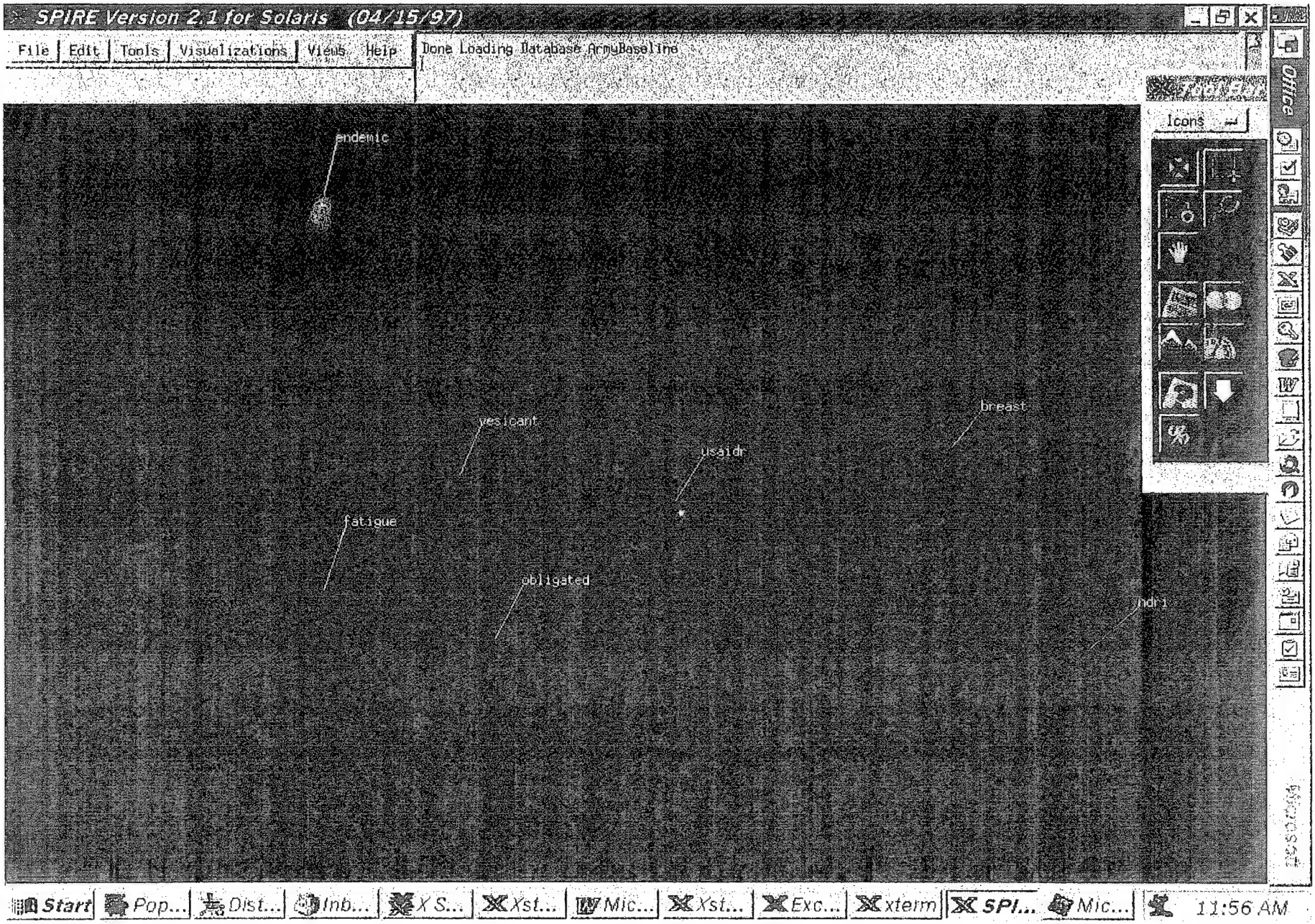
20 MRMC files

Prior to Processing into Document Subsets

MRMCBaseline-G1.jpg

MRMCBaseline-T1.jpg


SPIRE Version 2.1 for Solaris (04/15/97)
File
Edit
Tools
Visualizations
Views
Help
Done Loading Database ArmyBaseline
Icons
Start
11:58 AM


SPIRE Version 2.1 for Solaris (04/15/97)
File
Edit
Tools
Visualizations
Views
Help
Done Loading Database ArmyBaseline
endemic
vesicant
usaidr
breast
fatigue
obligated
hdr1
Icons
Start
Pop...
Dist...
Inb...
X S...
Xst...
Mic...
Xst...
Exc...
xterm
SPI...
Mic...
11:56 AM

Attachment 4

Separate Galaxies and Themescape Visualizations

of

ASBREM93, ASBREM95, and RAD94 Data


ASBREM93-G1.JPG

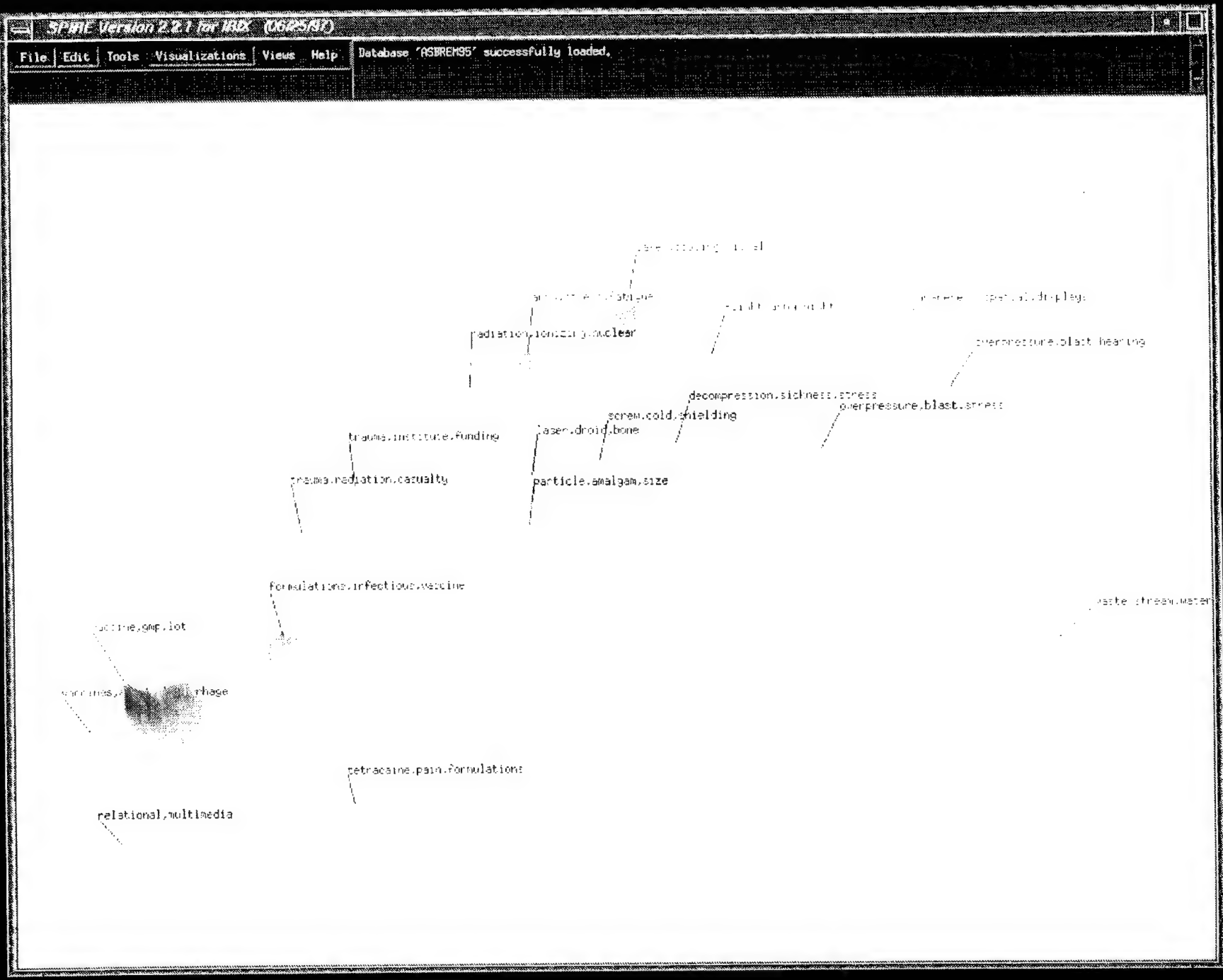

ASBREM93-T1.JPG


ASBREM95-G1.jpg

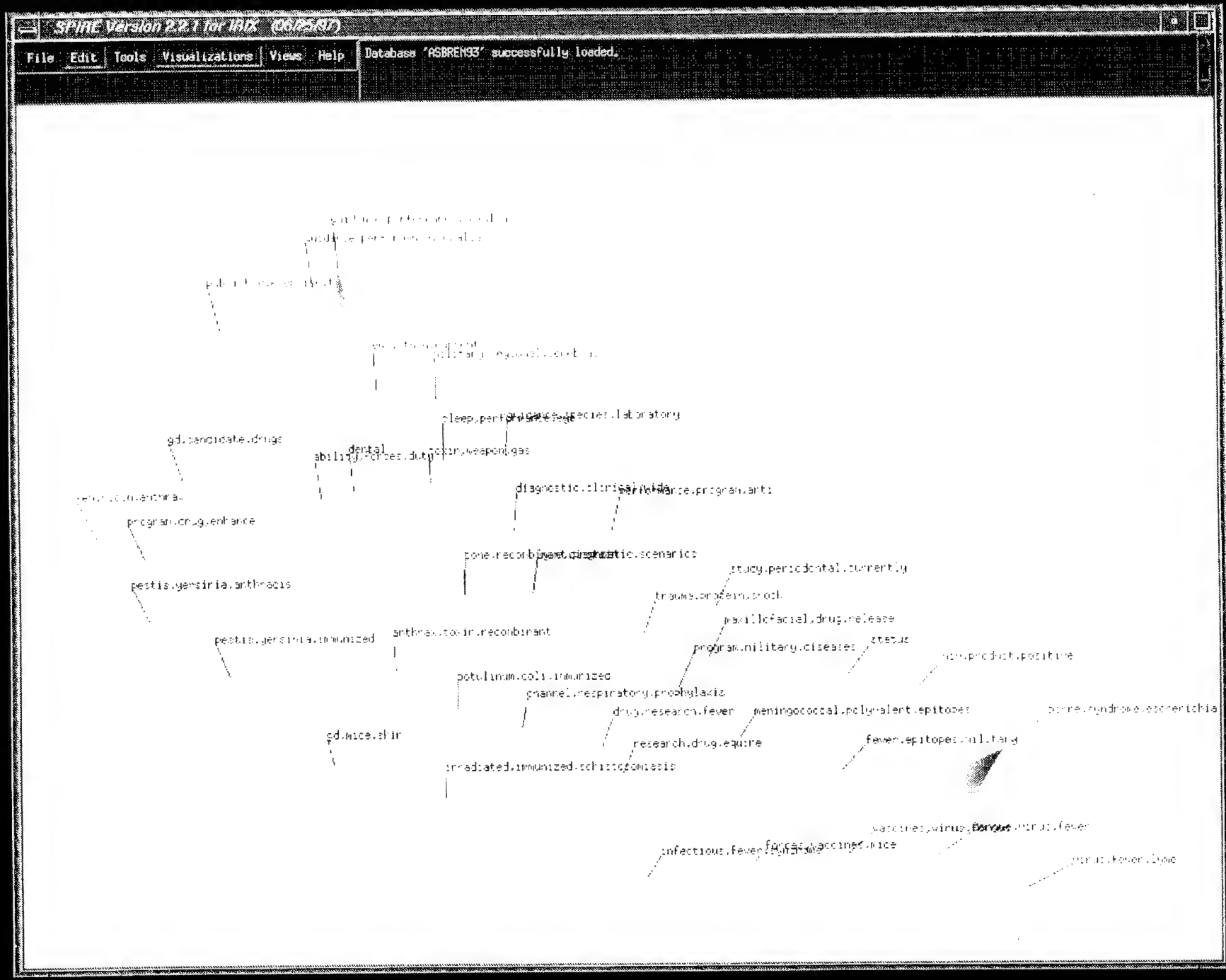

ASBREM95-T1.jpg

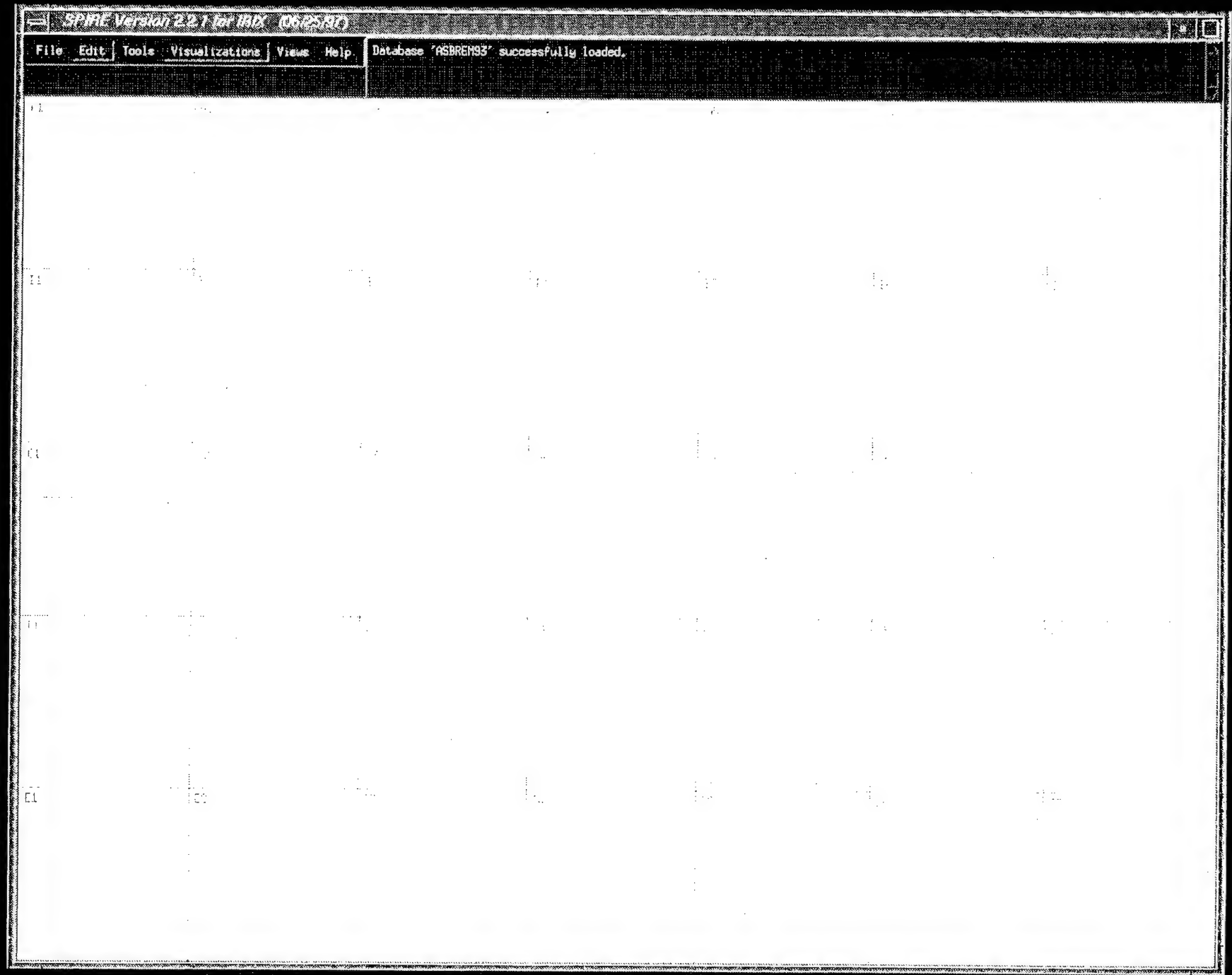

RAD-G1.jpg

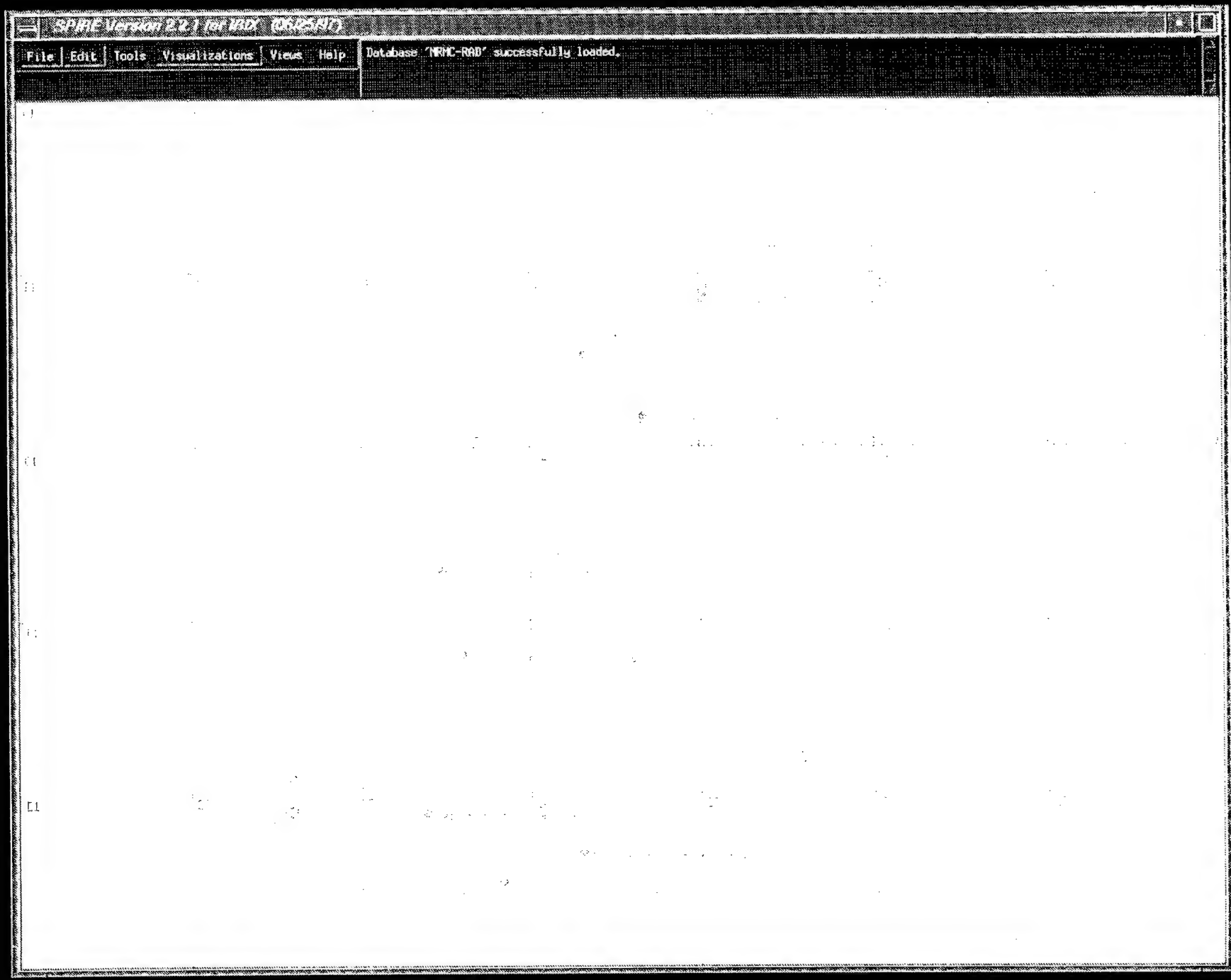

RAD-T1.jpg


SPIRE Version 2.2.1 for IRIX (O2/SAT)
File
Edit
Tools
Visualizations
Views
Help
Database 'ASBREM95' successfully loaded.
radiation.ionizing.nuclear
decompression.sickness.stress
overpressure.blast.hearing
screw.cold.shielding
trauma.institute.funding
laser.droid.bone
trauma.radiation.casualty
particle.amalgam.size
formulations.infectious.vaccine
vaccine.gmp.iot
tetracaine.pain.formulations
relational.multimedia


SPIRE Version 2.2.1 for IRIX (06.PSAT)
File
Edit
Tools
Visualizations
Views
Help
Database 'ASBREM95' successfully loaded.


SPINE Version 2.2.1 for IRIX (CBESRI)
File
Edit
Tools
Visualizations
Views
Help
Database 'ASBREM93' successfully loaded.
gd.candidate.drugs
dental
program.drug.enhance
pestis.yersiria.anthracis
pestis.yersiria.immunized
anthrax.toxin.recombinant
potulinum.coli.immunized
channel.respiratory.prophylaxis
drug.research.fever
meningococcal.polyvalent.epitopes
gd.mice.skin
research.drug.equire
irradiated.immunized.schistosomiasis
fever.epitopes.military
study.periodontal.currently
trauma.problem.oral
maxillofacial.drug.release
program.military.diseases
virus.product.positive
pone.syndrome.escherichia
infectious.fever
forces.vaccines.mice


SPIRE Version 2.2.1 for IRIX (06/25/96)
File
Edit
Tools
Visualizations
Views
Help
Database "ASBREM93" successfully loaded.


SPIRE Version 2.2.1 for VAX (MRMC-RAD)
File
Edit
Tools
Visualizations
Views
Help
Database 'MRMC-RAD' successfully loaded.


File
Edit
Tools
Visualizations
Views
Help
Database "MRMC-RAD" successfully loaded.
biological.exposure,countermeasures

Attachment 5

Combined Galaxies and Themescape Visualizations

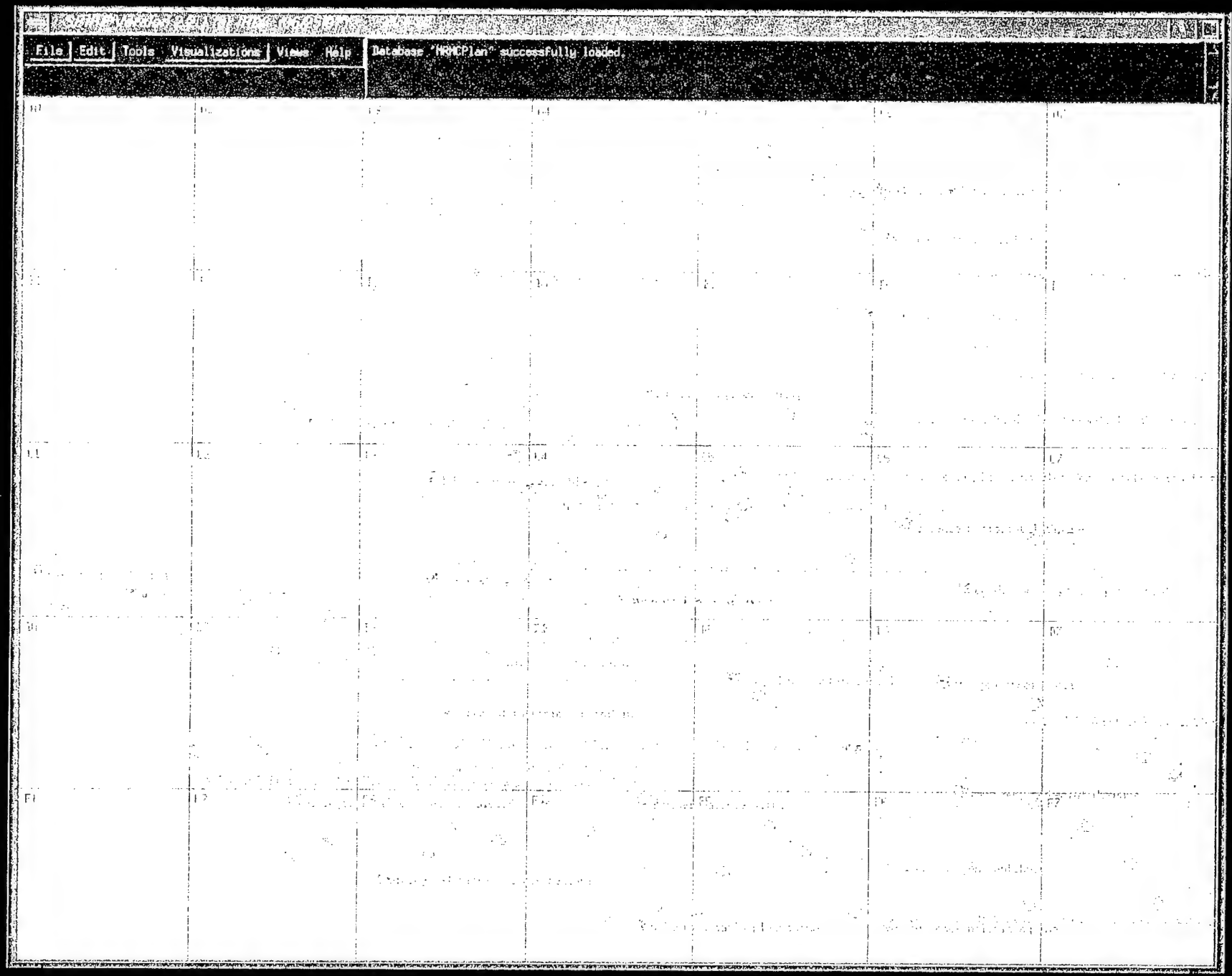
of

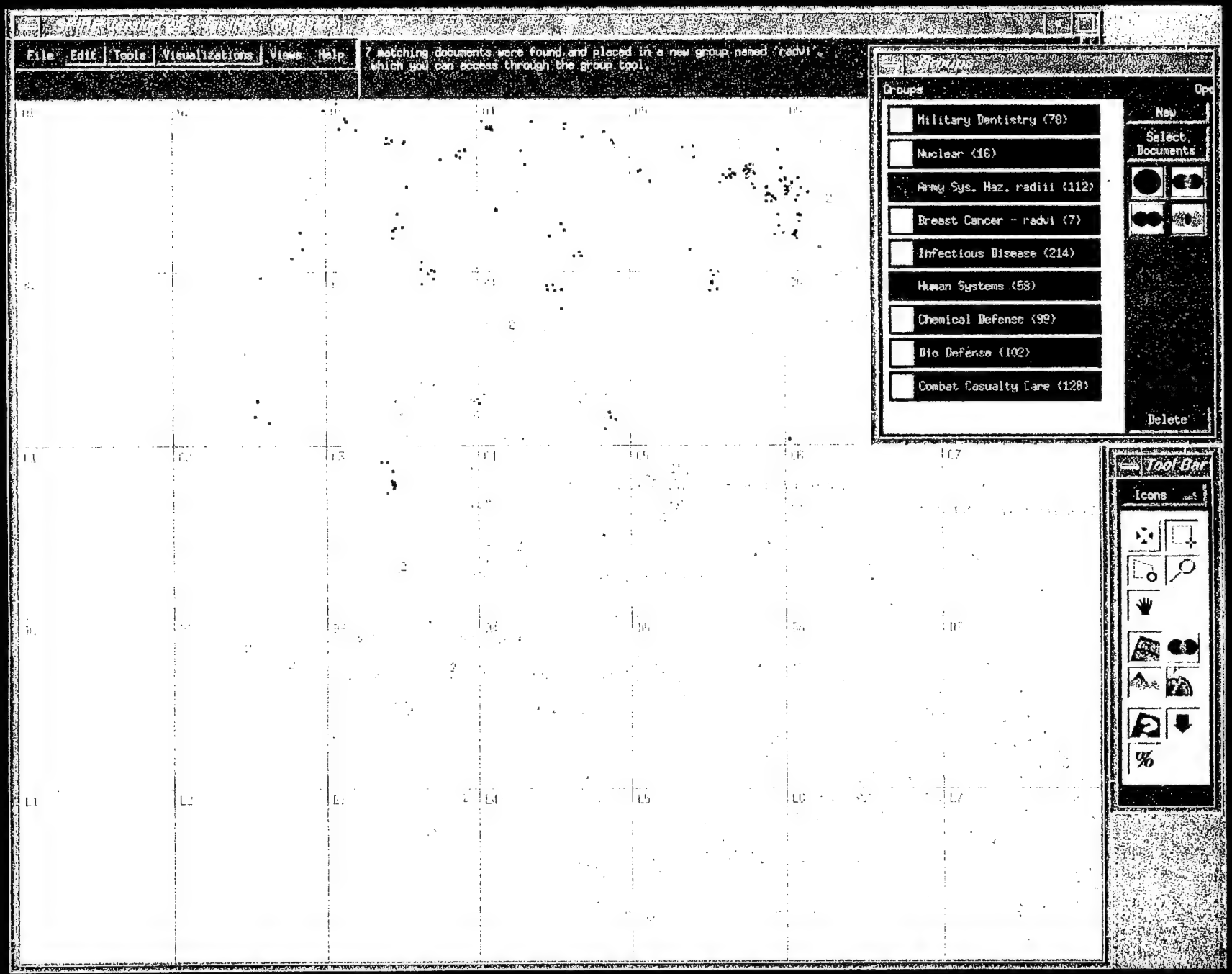
800 document segment files from MRMC Planning Documents,
and
21 segments from Army Biomedical Technical Area Plans.

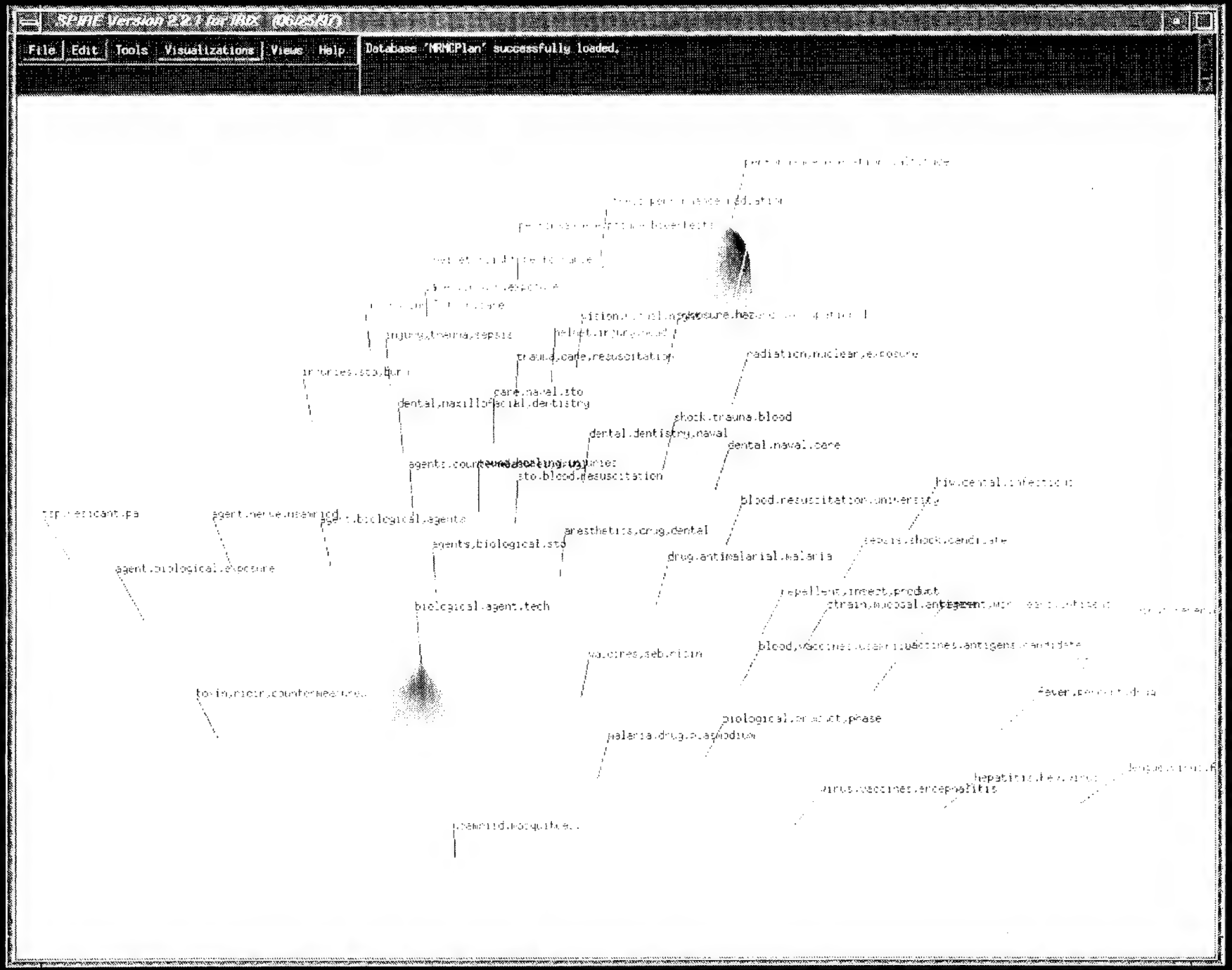

MRMCPlan-G1.jpg


MRMCPlan-Full.jpg


MRMCPlan-T1.jpg


File
Edit
Tools
Visualizations
Views
Help
Database "MRMCPlan" successfully loaded.


File
Edit
Tools
Visualizations
Views
Help
7 matching documents were found and placed in a new group named 'radvi', which you can access through the group tool.
Groups
Military Dentistry (78)
Nuclear (16)
Army Sys. Haz. radiii (112)
Breast Cancer - radvi (7)
Infectious Disease (214)
Human Systems (58)
Chemical Defense (99)
Bio Defense (102)
Combat Casualty Care (128)
New
Select Documents
Delete
Tool Bar
Icons


SPIRE Version 2.2.1 for IRIX (05/05/97)
File
Edit
Tools
Visualizations
Views
Help
Database 'MRMCPlan' successfully loaded.
injury,thermal,sepsis
trauma,care,resuscitation
radiation,nuclear,exposure
dental,maxillofacial,dentistry
shock,trauma,blood
dental,dentistry,naval
dental,naval,care
sto,blood,resuscitation
hiv,dental,infection
blood,resuscitation,university
agent,biological,exposure
anesthetics,drug,dental
agents,biological,sto
drug,antimalarial,malaria
repellent,insect,product
biological,agent,tech
vaccines,sebution
malaria,drug,plasmodium
virus,vaccines,encephalitis

Attachment 6

Time Slice Visualizations

of

MRMC Planning documents

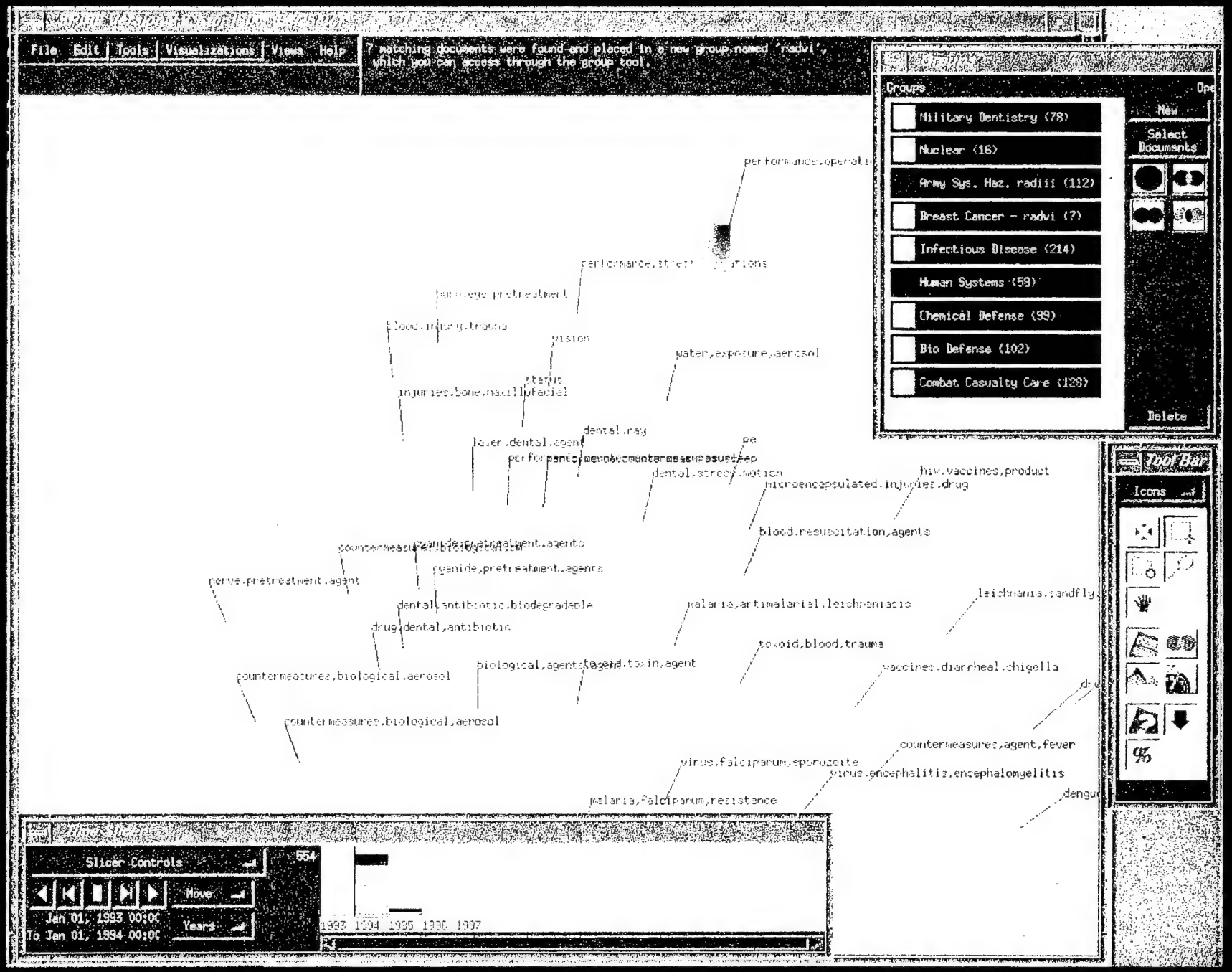
for

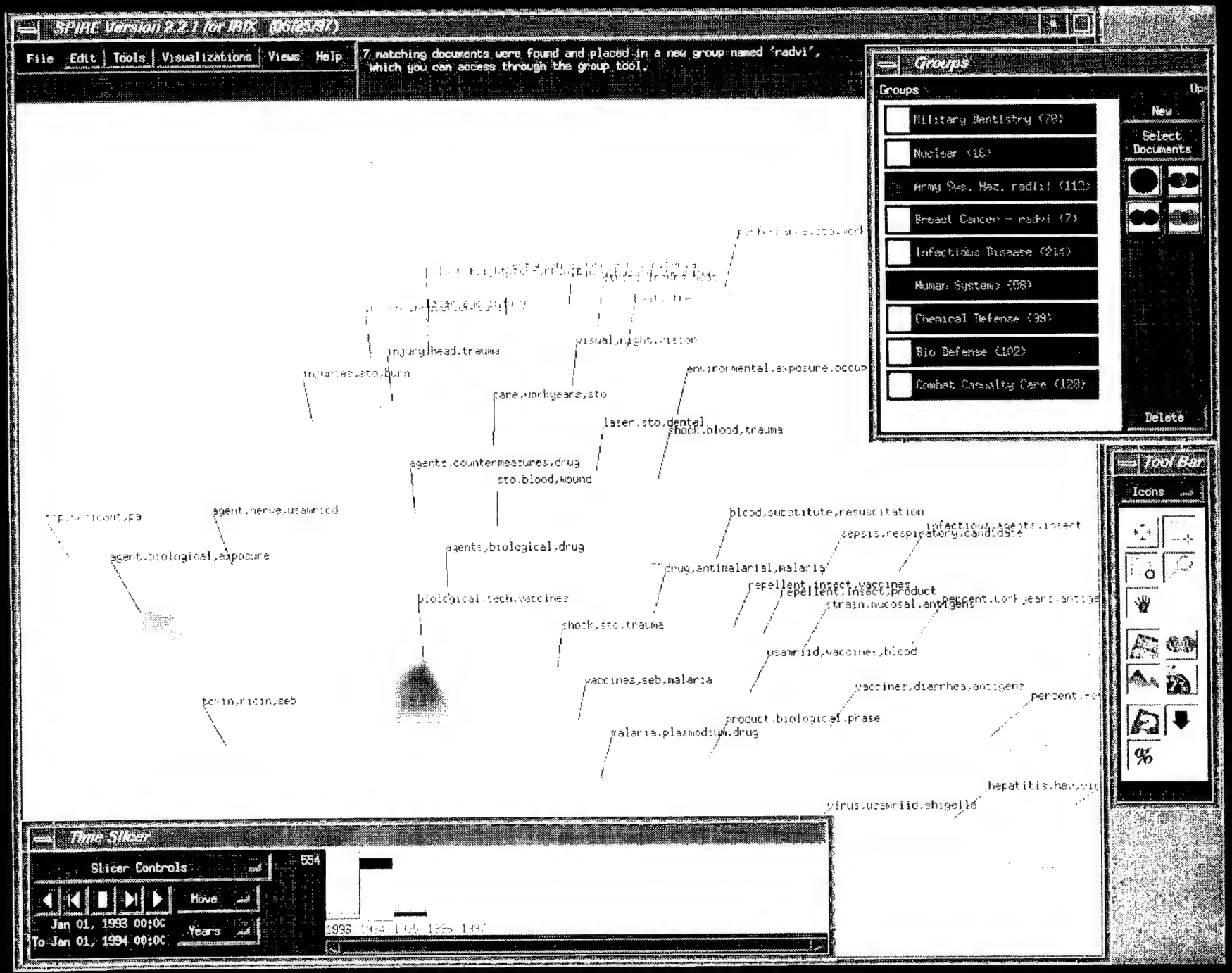
1993, 1994 and 1995

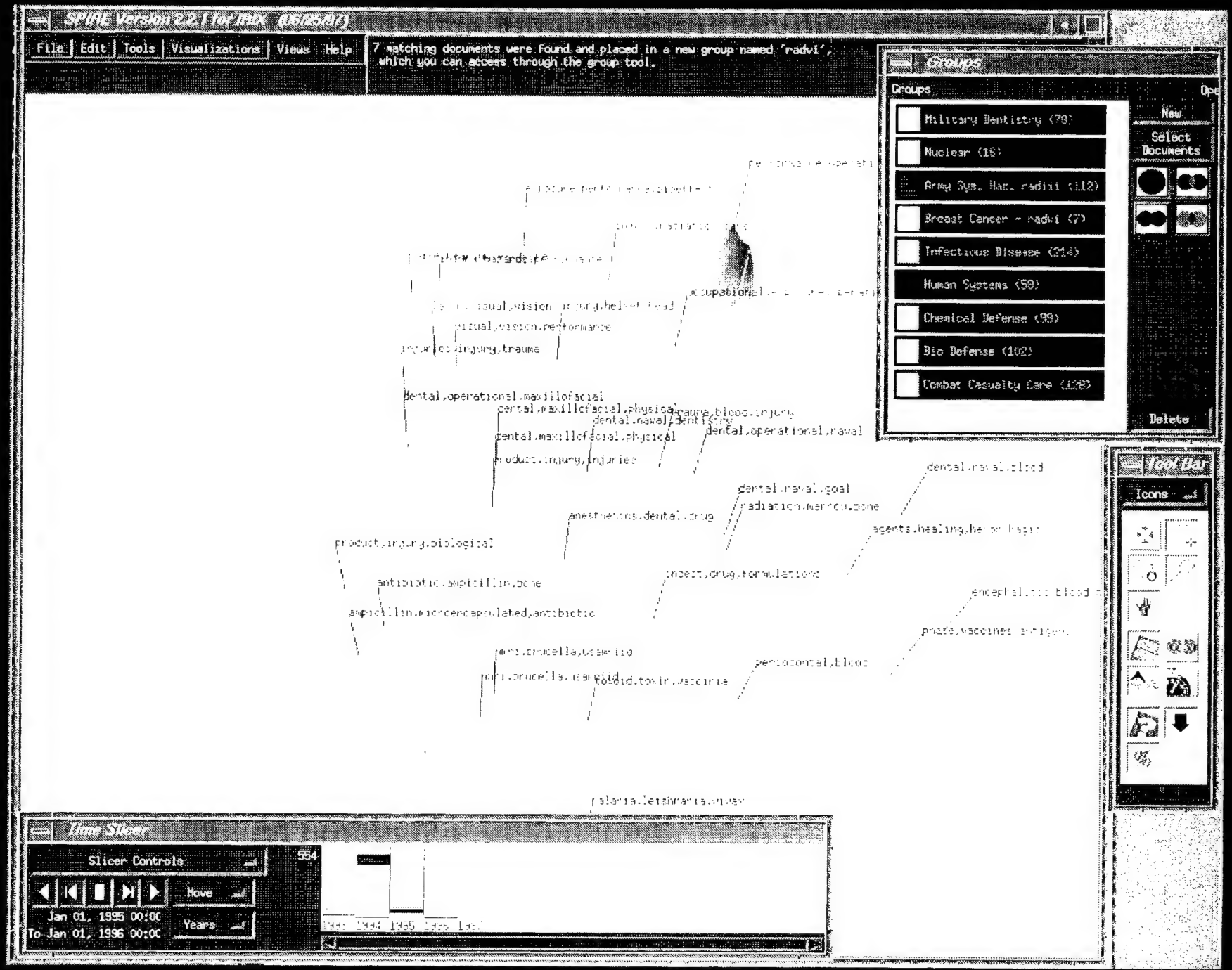

MRMCPlan-1993.jpg


MRMCPlan-1994.jpg


MRMCPlan-1995.jpg


File Edit Tools Visualizations Views Help
7 matching documents were found and placed in a new group named 'radvi', which you can access through the group tool.
Groups
Military Dentistry (78)
Nuclear (16)
Army Sys. Haz. radiii (112)
Breast Cancer - radvi (7)
Infectious Disease (214)
Human Systems (59)
Chemical Defense (99)
Bio Defense (102)
Combat Casualty Care (128)
New
Select Documents
Delete
Tool Bar
Icons
performance,operation
performance,stress
airways,pretreatment
blood,injury,trauma
vision
water,exposure,aerosol
status
injuries,bone,maxillofacial
dental,ray
lower,dental,agent
dental,stress,motion
hiv,vaccines,product
microencapsulated,injuries,drug
blood,resuscitation,agents
countermeasures,biological,agents
cyanide,pretreatment,agents
nerve,pretreatment,agent
dental,antibiotic,biodegradable
drug,dental,antibiotic
malaria,antimalarial,leishmaniacis
leishmania,sandfly
toxoid,blood,trauma
biological,agents,toxoid,toxin,agent
vaccines,diarrheal,shigella
countermeasures,biological,aerosol
countermeasures,biological,aerosol
countermeasures,agent,fever
virus,falciparum,sporozoite
virus,encephalitis,encephalomyelitis
malaria,falciparum,resistance
dengue
Slicer Controls
654
Move
Jan 01, 1993 00:00
To Jan 01, 1994 00:00
Years
1993 1994 1995 1996 1997


SPIRE Version 2.2.1 for IRIX (6.5-R7)
File Edit Tools Visualizations Views Help
7 matching documents were found and placed in a new group named 'radvi', which you can access through the group tool.
Groups
Groups
Military Dentistry (78)
Nuclear (16)
Army Sys. Haz. radii (113)
Breast Cancer - radvi (7)
Infectious Disease (214)
Human Systems (58)
Chemical Defense (98)
Bio Defense (102)
Combat Casualty Care (128)
New
Select Documents
Delete
Tool Bar
Icons
injury,head,trauma
injuries,sto,burn
visual,night,vision
environmental,exposure,occup
care,workyears,sto
laser,sto,dental
shock,blood,trauma
agents,countermeasures,drug
sto,blood,wound
agent,nerve,usamriid
agent,biological,exposure
agents,biological,drug
blood,substitute,resuscitation
sepsis,respiratory,candidate
drug,antimalarial,malaria
repellent,insect,vaccines
repellent,insect,product
biological,tech,vaccines
shock,sto,trauma
usamriid,vaccines,blood
vaccines,seb,malaria
vaccines,diarrhea,antigens
toxin,ricin,seb
product,biological,phase
malaria,plasmodium,drug
hepatitis,hev
virus,usamriid,shigella
Time Slicer
Slicer Controls
554
Move
Jan 01, 1993 00:00
To Jan 01, 1994 00:00
Years
1993


SPINE Version 2.2 (for IRIX (R6/64))
File
Edit
Tools
Visualizations
Views
Help
7 matching documents were found and placed in a new group named 'radvi', which you can access through the group tool.
Groups
Groups
Military Dentistry (78)
Nuclear (15)
Army Sym. Max. radii (112)
Breast Cancer - radvi (7)
Infectious Disease (214)
Human Systems (50)
Chemical Defense (99)
Bio Defense (102)
Combat Casualty Care (128)
New
Select Documents
Delete
Tool Bar
Icons
Time Slicer
Slicer Controls
554
Move
Jan 01, 1995 00:00
To Jan 01, 1996 00:00
Years
1965
1966
dental,operational,maxillofacial
dental,maxillofacial,physical
dental,naval,dentistry
dental,maxillofacial,physical
product,injury,injuries
dental,operational,naval
dental,naval,goal
radiation,marrow,bone
anesthetics,dental,drug
product,injury,biological
antibiotic,ampicillin,bone
ampicillin,microencapsulated,antibiotic
insect,drug,formulations
dental,naval,blood
agents,healing,bone,tissue
encephalitis,blood
polio,vaccines,antigen
periodontal,blood
toxoid,toxin,vaccine
malaria,leishmania,virus

APPENDIX C

VISUALIZATION OF MEDICAL SENSOR DATA:

SIGNAL PROCESSING CONCEPT EXPLORATION

PNWD-2412

Visualization of Medical Sensor Data:
Signal Processing Concept Exploration

N. E. Miller
P. D. Whitney
D. S. Daly

December 1997

Prepared for
Walter Reed Army Institute of Research Division of Surgery
by the Pacific Northwest National Laboratory
under U.S. Army Medical Research and Material Command
Contract 26186A

DISCLAIMER

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC06-76RLO 1830

PNWD-2412

Visualization of Medical Sensor Data:
Signal Processing Concept Exploration

N. E. Miller
P. D. Whitney
D. S. Daly

December 1997

Prepared for
Walter Reed Army Institute of Research Division of Surgery
under U.S. Army Medical Research and Material Command Contract 26186A

Pacific Northwest National Laboratory
Richland, WA 99352

Abstract

The U.S. Army Medical Research and Material Command (MRMC) has commissioned the Pacific Northwest National Laboratory (PNNL) to work with the Walter Reed Army Institute of Research Division of Surgery (WRAIR Surgery Division) to visualize medical sensor data. The WRAIR Surgery Division has a considerable mass of data on patients in various conditions. The **Signal Processing Concept Exploration** Task was undertaken as part of a larger project to explore the concept of applying Spatial Paradigm for Information Retrieval and Exploration (SPIRE™) visualization technologies to several areas of medicine. The Signal Processing Concept Exploration Task was a proof-of-concept task. This is the final report for that task.

We have applied the SPIRE analysis paradigm to highly sampled multiple sensor medical data. The data for our analysis were from three patients who had undergone open-heart surgery at WRAIR. Each patient was post-operatively monitored at 1000 Hz for several hours with EKG, Direct Arterial Pressure, and Pulse Blood Oximeter sensors, each recording simultaneously. Our analysis of the medical sensor data is based on the definition of an "event" as the activity associated with a particular heartbeat for a particular patient. We used a custom pattern-matching technique to detect distinct heartbeats within the data for each sensor.

In SPIRE, the analysis is based on thematic similarity between documents. Documents thus serve as the core objects within the analysis. A text engine automatically extracts the best word features for the entire document collection. These best word features together with the word frequency counts and word associations are then used to create a numerical signal for each document object. This high-dimensional numerical signal contains the "attributes" for the document objects. These "attributes" are used to cluster the documents and then to project down to a 2-dimensional view called Galaxy, which is based on a Principal Components Analysis (PCA).

The SPIRE text-to-signal generator, which converts words within unstructured text to mathematical signals, does not operate on numerical data such as medical sensor data. Thus, we wrote a custom feature selection engine to extract the features in the neighborhood for each heartbeat. We present separate Principal Component-Based visualizations for each patient based on a different combination of initial variables. A videotape that shows 3-D animated visualizations for the various patient/variable sets accompanies this report.

This report contains several SPIRE visualizations of a medical corpus comprised of Medline abstracts with publication dates between 1994 and spring 1997. This data set was built for this task. In the summer of 1997, this corpus and the SPIRE source code were installed on "medic1" at the WRAIR Surgery Division.

This report also contains several suggestions for follow-on work. "MediSense" is proposed to move this proof-of-concept work to the state of a prototype that would allow near-real-time access to this type of exploratory data analysis for a single patient and perform the integration and visualizations across patients.

Acknowledgments

The authors gratefully acknowledge the expert assistance of our technical editor, Sharon Eaton, whose efforts are also included in this final report. We would also like to acknowledge Dennis McQuerry, who performed the Medline queries and constructed the corpus of Medline abstracts that were used in the visualizations found in Appendix A.

Contents

APPENDIXES

FIGURES

1.0 Introduction

The U.S. Army Medical Research and Material Command (MRMC) has commissioned the Pacific Northwest National Laboratory (PNNL) to work with the Walter Reed Army Institute of Research Division of Surgery (WRAIR Surgery Division) to visualize medical sensor data. The WRAIR Surgery Division has a considerable mass of data on patients in various conditions and needs to be able to analyze this data and search for correlations among various data types.

The work performed consisted of four major tasks. This report focuses on the third task, **Signal Processing Concept Exploration**, which was undertaken to explore the concept of applying visualization technologies to several fields of medicine and resulted in the definition of three activities.

1.1 Original Scope of Work and Deliverables

Three activities comprised the original statement of work. The scope for these activities as stated in the original statement of work and their corresponding deliverables are specified below.

1.1.1 Original Scope of Work

Three activities comprised the original scope of work for this contract:

- Activity A - Research options for managing and correlating information recorded from multiple sensors from multiple patients. Research the feasibility of entering this information into a database and look for similarities in a) frequency spectrum, b) transients (time domain response), and c) power/intensities among sensors on the same person and among different people according to their physical condition. Research the ability to tag the sensor information with the diagnosed illness and the symptoms and then further correlate these with the sensor inputs and the correlation among inputs.

- Activity B - Research the feasibility of, and the options for, using the database created in Activity A as a tool for recognition of symptoms. For example, the client would like to input sensor data to get a visual depiction of other patients with a) similar symptoms, b) similar sensor suites, and c) similar conditions or treatment strategies.

- Activity C - Scope the level of effort to establish a medical document database for researchers and to apply the Spatial Paradigm for Information Retrieval and Exploration (SPIRE) technology to process and visualize the contents of the database. Identify technical papers and articles from key journals. Process a test data set. Provide recommendations on database design and construction.

1.1.2 Deliverables

- Activities A and B - Report on research finding and technology assessment. Where feasible, demonstrate technology options

- Task C - Present data set analysis demonstrating the results live on SGI system. Include hard copies of analysis visualizations.

1.2 Modified Scope of Work

The original scope of the Signal Processing Concept Exploration Task was modified over the course of the project through telephone conversations and two site visits with Dr. Frederick Pearce of WRAIR Surgery Division and his staff. Several factors elicited this modified scope of work: the number of patients was small (three) and the ancillary data required correlate patient symptoms and diagnosis were not available. We did, however, receive three - four

hours of sensor data for each patient that allowed us to very successfully address the fundamental aspects of extending the SPIRE analysis paradigm to the collection of sensor data for each patient individually.

1.2.1 Movement Away from Integrating Work into SPIRE

The intent of this research was to apply the SPIRE analysis paradigm to highly sampled multiple sensor medical data. The Signal Processing Concept Exploration Task was a proof-of-concept task; it was not a task to deliver a prototype. The SPIRE text-to-signal generator, which converts words within unstructured text to mathematical signals, does not operate on numerical data such as medical sensor data. Thus, we wrote a feature selection engine in S-PLUS® (Statistical Sciences, Inc.) — a quick statistical prototyping language — and performed all the feature extraction processing in S-PLUS. The dimensionality reduction analysis was performed in MATLAB® (The Mathworks, Inc.). Because the interaction tools for SPIRE are specific to text, we used Data Desk® (Data Description, Inc.) to perform the visualizations contained in Section 4.0, "Visualizations of Analysis Results."

1.2.2 Report of Other Activities

Several other activities that were not explicitly outlined in the original statement of scope were undertaken by the Signal Processing Concept Exploration Task. These included delivery of the SPIRE analysis tool, some custom Perl scripts, and a custom Medline Abstract data set.

1.2.2.1 Delivery of SPIRE, Perl Scripts Document Tools

In late spring 1997 during a site visit to Walter Reed, we installed the SPIRE software on the client's SGI system, medic1. Subsequent phone conversations with David Scott resulted in the creation of some custom software tools (Perl scripts) to get various medical documents into a SPIRE-friendly format. These Perl scripts were installed on medic1.

1.2.2.2 Creation of Medline Corpora

A deliverable associated with Activity C was to visualize a Medical Data Set using SPIRE. To perform this function, we first built a medical corpus. The corpus we constructed was comprised of Medline abstracts from the National Library of Medicine (NLM) with publication dates between 1994 and spring 1997. The harvested abstracts were the result of a query on "hemorrhage" or "shock." Approximately 16,000 titles were returned from this query. Various SPIRE visualizations are included in Appendix A, "Medline Corpus Visualizations," from this data set. This corpus was also installed on medic1. Costs associated with purchasing the collection of CDs containing Medline abstracts are also available in Appendix A. However, the NLM now makes the Medline database available over the network at no cost. If the user interface for this harvesting process is now functional, then it is likely not necessary to buy the CDs.

2.0 Data for Analysis

We received several shipments of data for this project. In April 1997, we received data for Patient B. A month later, we received a new file for Patient B and data for Patient F. In July 1997, we received the final analysis data sets. These included a reprocessing of Patient B and data for two new patients, Patients H and W. At this time, we were instructed not to proceed with patient F. The results contained in this paper pertain to the final data set: Patients B, H, and W.

2.1 Data Preparation - Codas Reformatting

The binary files that were provided were in a proprietary format known as CODAS. We wrote a FORTRAN program to read these binary files and output patient-specific ASCII files for the three sensor fields of interest. These ASCII files were read by our custom S-PLUS software.

2.2 Final Datasets

Each of the patients, B, H, and W, had undergone open-heart surgery at WRAIR and were post-operatively monitored for several hours with sensors recording simultaneously. Some sensors were non-invasive and would likely be available on a battlefield, while other invasive measurements would likely only be available in a hospital setting. We were provided three types of sensor measurements for each patient:

- EKG (non-invasive)
- Direct Arterial Pressure (invasive)
- Pulse Blood Oximeter (non-invasive).

During the time of monitoring, the patient presumably received various different treatment strategies. The details of these treatments were unknown to us, as were the times of the various treatments. We had no personal information about any patient such as age, gender, etc., except that one of the patients, H, did have a pacemaker.

The data were recorded at 1000 Hertz. The following table summarizes the length of observation for each patient:

Table 1 - Patient Observation Period

	Number of Records	Number of Hours
Patient B	10752001	2.99
Patient H	14005124	3.89
Patient W	14751050	4.10

Three or four hours of data were measured for each patient. Not all records were usable, as any of the three sensors could potentially have bad data.

3.0 Analysis Approach

In SPIRE, the analysis is based on thematic similarity among documents. Documents thus serve as the core objects within the analysis. A text engine called SID is used to automatically extract the best word features for the entire document collection. These best word features together with the word frequency counts and word associations are then used to create a numerical signal for each document object. This high-dimensional numerical signal contains the "attributes" for the document objects. These "attributes" are used to cluster the documents and then to project down to a 2-dimensional view called Galaxy, which is based on a Principal Components Analysis (PCA) (see Section 3.6.1, "Theory - Principal Components Analysis").

Our analysis of the medical sensor data is based on the definition of an "event" as the activity associated with a particular heartbeat for a particular patient. The three sensors take measurements synchronously in time. We process the time series for each sensor individually to identify the onset time for each heartbeat. The onset times across sensors are then organized by the specific heartbeat to which they refer. These windows around each heartbeat are used to define a neighborhood within which other calculations will be made for a given sensor or across sensors that will ultimately be used to calculate features for our analysis. The features that we calculate for each heartbeat-to-heartbeat event were predefined for the most part by Dr. Pearce and are discussed in more detail in Section 3.2, "Time Picks for Feature Definitions." These values are then organized into a feature matrix where each row represents a distinct consecutive heartbeat. Dimensionality reduction is then accomplished on this feature matrix using a PCA. The events (which we can think of as heartbeats) are then visualized in 2- or 3-space using the PCA scores.

3.1 Definition of Events

The first 4000 data records for Patient B are illustrated in Figure 1. The vertical lines indicate the time at which a heartbeat is detected for the three sensors. Notice that the sensors lag each other in time. The top EKG panel shows the first five events; the middle and bottom panels show the first four events for the A-line and Pulse Blood Oximeter data respectively.

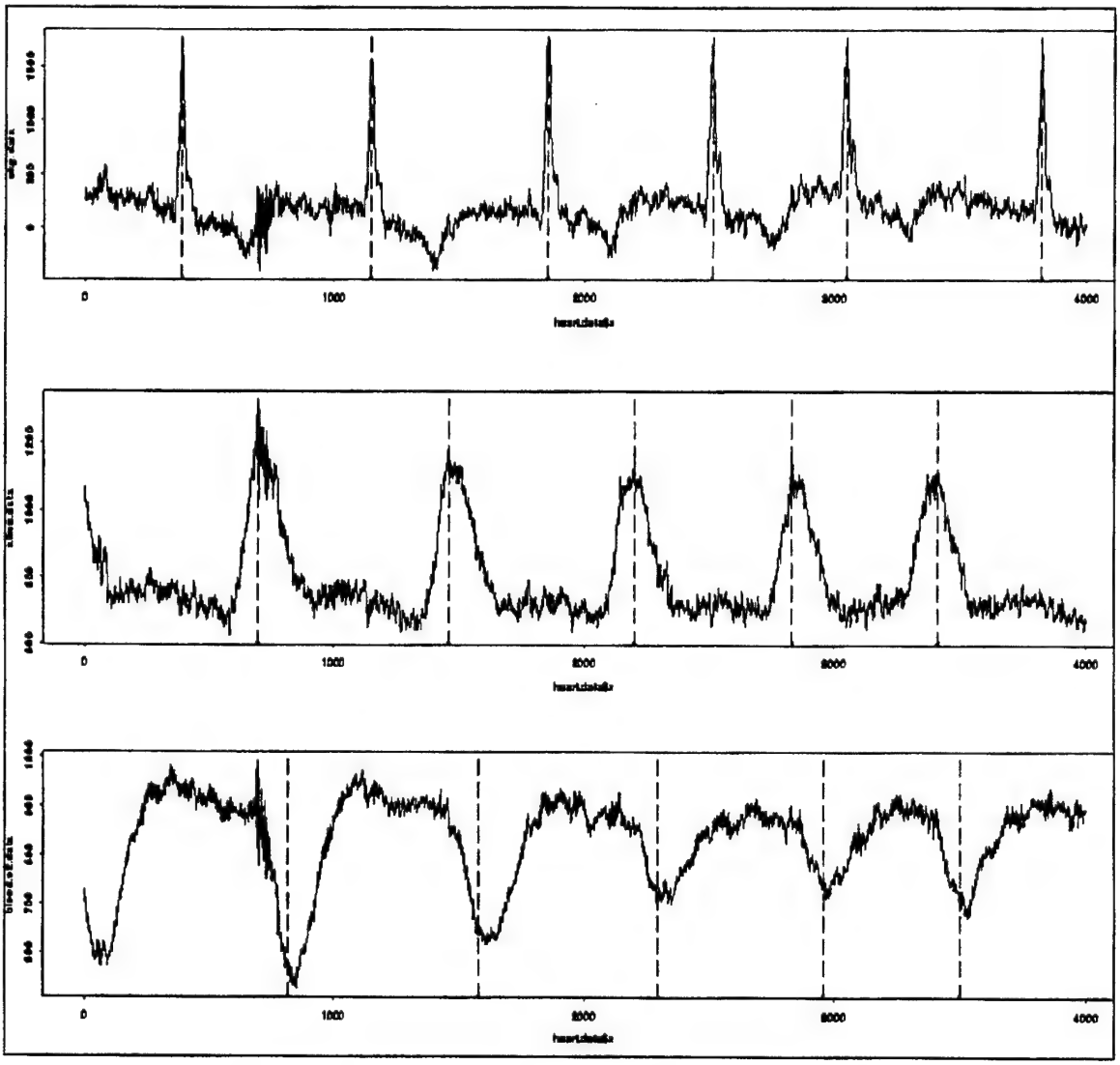


Figure 1 - Patient B Data - Detection of Heartbeat Events

Note that the A-line heartbeat lags the EKG heartbeat signal in time and that the Pulse Blood Oximeter heartbeat signal lags the A-line signal.

In order for the features from a given heartbeat to progress forward to the final analysis, all three sensors must have had all features identified and calculated. Although other alternatives to this approach, such as filling in for missing data exist — e.g., an EM algorithm (Little 1987) — we have kept things very simple for this research.

3.1.1 Pattern Matching Approach

To detect the time of the heartbeat in each of the three sensors for a given patient, an archetypical waveform was selected for each patient/sensor. The archetype contained the time of the maximum amplitude (or minimum amplitude for the Pulse Blood Oximeter sensor) heartbeat in the center of the waveform. For each patient/sensor time series, this archetype was moved across a window for the time series. Correlations were calculated between the underlying window and the archetype. Parameters allowed this setting to be changed. The value we used for all patient/sensor combinations was .90 with good results.

3.2 Time Picks for Feature Definitions

During our meeting with Dr. Frederick Pearce, several classes of features were discussed. These classes included time-domain-based features: simple heartbeat features, statistical summaries of simple heartbeat features, and frequency-domain-based features.

The majority of the time-domain-based features began with the identification of local maxima and minima for the smoothed sensor data. In row 1 of Figure 2, a small subset of the original sensor data for Patient B is shown as dots for the EKG, A-line, and Pulse Blood Oximeter data. The first and second derivatives are displayed in rows 2 and 3, respectively. The smoothed data are shown by the black curve. The smooth of the data was performed using a de Boor's cubic spline fit (de Boor 1978).

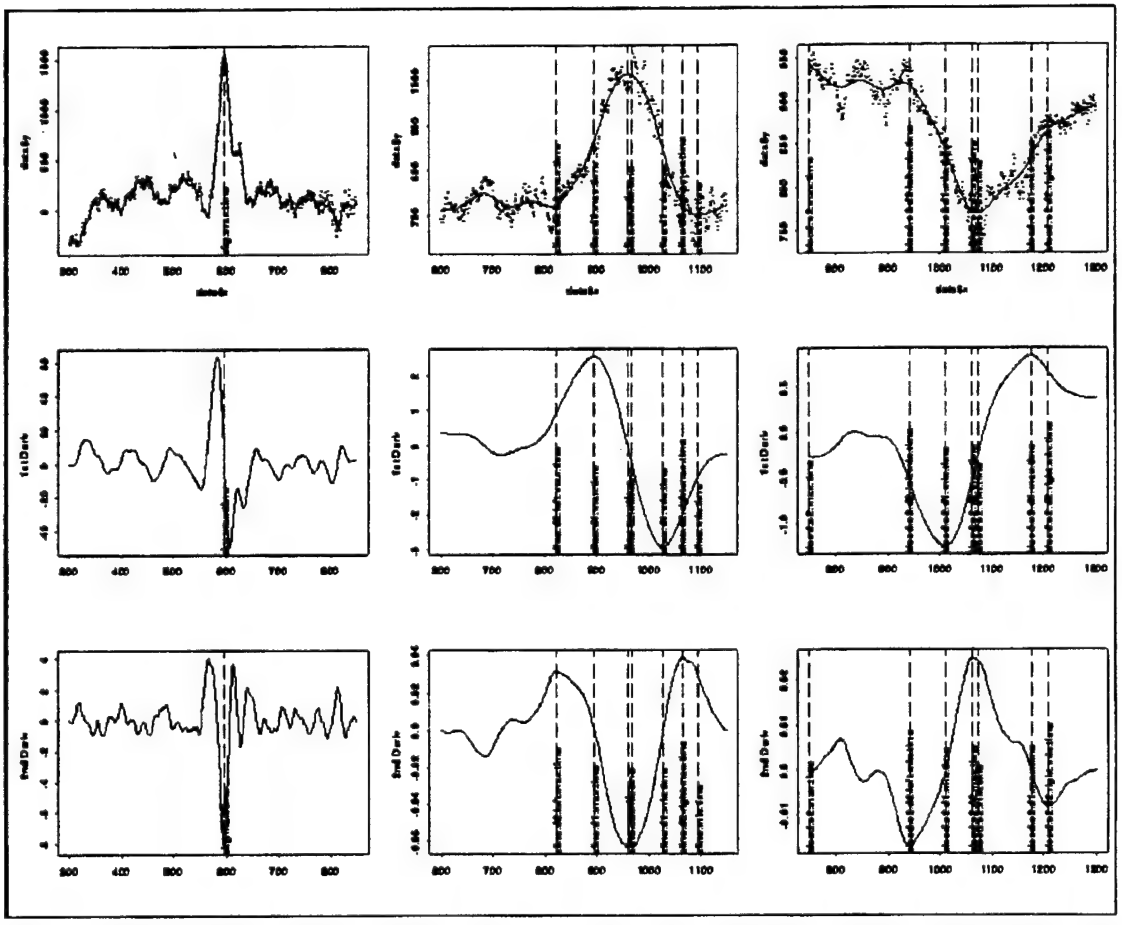


Figure 2 - Patient B Data - Sensor Values (columns) and Smooths (rows) Surrounding a Single Event with Associated Time Picks

The following times were identified for the smooths from each patient/heartbeat/sensor collection as illustrated in Figure 2:

- EKG maximum time
- A-line features (the order of the labels is as they occur in Figure 2 from left to right)
    1. second derivative left maximum time
    2. first derivative maximum time
    3. maximum time

4. second derivative minimum time
5. first derivative minimum time
6. second derivative right maximum time
7. minimum time

- Pulse Blood Oximeter features (the order of the labels is as they occur in Figure 2 from left to right.)
  1. maximum time (this was not used as arbitrary)
  2. second derivative left minimum time
  3. first derivative minimum time
  4. second derivative maximum time
  5. minimum time
  6. first derivative maximum time
  7. second derivative right maximum time

3.3 Simple Heartbeat Features

The time picks above were used as input to calculate many of the simple heartbeat features defined in Section 3.3.1, "Single Sensor Features," for sensor data used alone. Several features, as discussed in Section 3.3.2, "Multiple Sensor Features," were calculated by parameterizing the "donut" formed when the A-line data and the Pulse Blood Oximeter data for a given heartbeat interval were plotted against one another. In Section 3.3.3, "Simple Heartbeat Features," we look at the pattern of missing values in the data for Patients B, H, and W. The complete list of variables calculated in the analysis may be found in Appendix B, "Definition of Variable Names."

3.3.1 Single Sensor Features

The simple heartbeat features that were calculated using the data from a single sensor include those shown in the list below. For the "*diff" variables, reasonable assumptions were made about pulse rates based on past history in the time series. It is very helpful to refer back to Figure 2 when reading the definitions of the features.

Features Based on EKG

- ekg diff - *delta time* between ekg.max.time and previous ekg.max.time
- aline diff - *delta time* between aline.max.time and previous aline.max.time
- blood o2 diff - *delta time* between blood.o2.min.time and previous blood.o2.min.time
- R to Aline - *delta time* between ekg.max.time and aline.max.time for this heartbeat (fth)
- R to Oxi - *delta time* between ekg.max.time and blood.o2.min.time fth
- peak R - *amplitude* value at ekg.max.time fth

Features Based on A-line

- Aline amplitude - *delta amplitude*; amplitude value for aline.max.time minus amplitude value for aline.min.time
- peak Aline - *amplitude* value for aline.max.time fth
- high d1 Aline - value of first derivative or *slope* at aline.d1.max.time fth
- low d1 Aline - value of first derivative or *slope* at aline.d1.min.time fth
- d1 width Aline - *delta time* between aline.d1.min.time and aline.d1.max.time fth
- d2 width Aline - *delta time* between aline.d2.right.max.time and aline.d2.left.max.time
- high d2 left Aline - *value of second derivative* at aline.d2.left.max.time
- high d2 right Aline - *value of second derivative* at aline.d2.right.max.time
- low d2 Aline - *value of second derivative* at aline.d2.min.time
- integral Aline - area bounded by the A-line curve and the line segment determined by aline.d2.left.max.time and aline.d2.right.max.time

Features Based on Pulse Oximeter

- Oxi amplitude - *delta amplitude*; amplitude value for blood.ox.max.time minus amplitude value for blood.ox.min.time
- trough Oxi - *amplitude* value for blood.ox.min.time fth
- high d1 Oxi - value of first derivative or *slope* at blood.ox.d1.max.time fth
- low d1 Oxi - value of first derivative or *slope* at blood.ox.d1.min.time fth
- d1 width Oxi - *delta time* between blood.ox.d1.min.time and blood.ox.d1.max.time fth
- d2 width Oxi - *delta time* between blood.ox.d2.right.min.time and blood.ox.d2.left. max.time fth
- low d2 left Oxi - *value of second derivative* at blood.ox.d2.left.min.time
- low d2 right Oxi - *value of second derivative* at blood.ox.d2.right.min.time
- high d2 Oxi - *value of second derivative* at blood.ox.d2.max.time
- integral Oximeter - area bounded by the Pulse Blood Oximeter curve and the line segment determined by blood.ox.d2.left.min.time and blood.ox.d2.right.min.time

3.3.2 Multiple Sensor Features

Another set of 21 features was defined by parameterizing the "donut" shape achieved when the A-line data are plotted against the Pulse Blood Oximeter data between each heartbeat as shown in Figure 3.

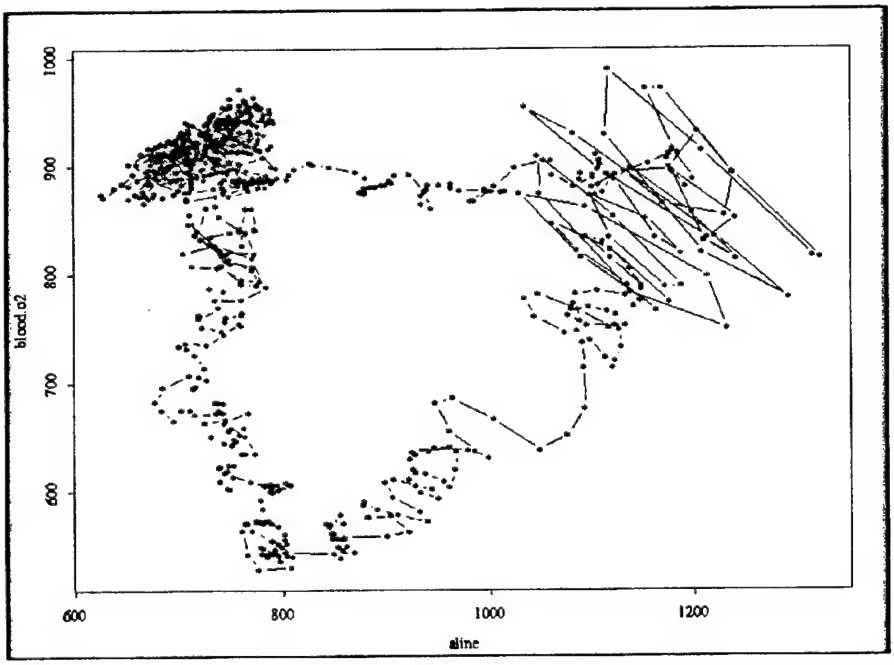


Figure 3 - Patient B - Pulse Blood Oximeter versus A-line Data, the "Donut"

Mathematically, we track the "donut" by using the coefficients from the parametric regression b-splines fit of the A-line and Pulse Blood Oximeter data to the same temporal data. The fit was accomplished by first centering and scaling the temporal data. Then the temporal interval was broken into subintervals — in our case, 10. The spline is a collection of polynomials of degree less than or equal to three such that the second derivatives agree at the end of each subinterval and such that the second derivative is continuous. The list of these variables is show below:

Spline-based Features for A-line Data

- aline ns Intercept
- aline ns 1
- aline ns 2
- aline ns 3
- aline ns 4
- aline ns 5
- aline ns 6
- aline ns 7
- aline ns 8
- aline ns 9
- aline ns 10

Spline-based Features for Pulse Blood Oximeter Data

- blood o2 ns
- blood o2 ns 1
- blood o2 ns 2
- blood o2 ns 3
- blood o2 ns 4
- blood o2 ns 5
- blood o2 ns 6
- blood o2 ns 7
- blood o2 ns 8
- blood o2 ns 9
- blood o2 ns 10

3.3.3 Pattern of Missing Values

A graphical summary of all missing data was performed for each of the three patients for the simple heartbeat features. These plots may be found in Appendix C, "Simple Features – Pattern of Missing Values." The variable numbering used in the plots follows the numbering scheme specified in Appendix B, "Definition of Variable Names." Note that each time the variable is missing, a dot is plotted over the three to four hours of the data record for each patient. Missing data can occur because the pattern recognizer fails or because of bad data. We have made no attempt to assign percentages to the causes of missing features.

In order for a set of features from a given heartbeat to be retained for analysis, all features had to be present. We defined a feature as missing if and only if we had located the heartbeat in the EKG data but were then unable to calculate some subsequent feature from the list shown in Appendix B.

Figure 4 shows a span of data from Patient H where the pattern matcher failed to pick out the EKG data because of an inverted peak.

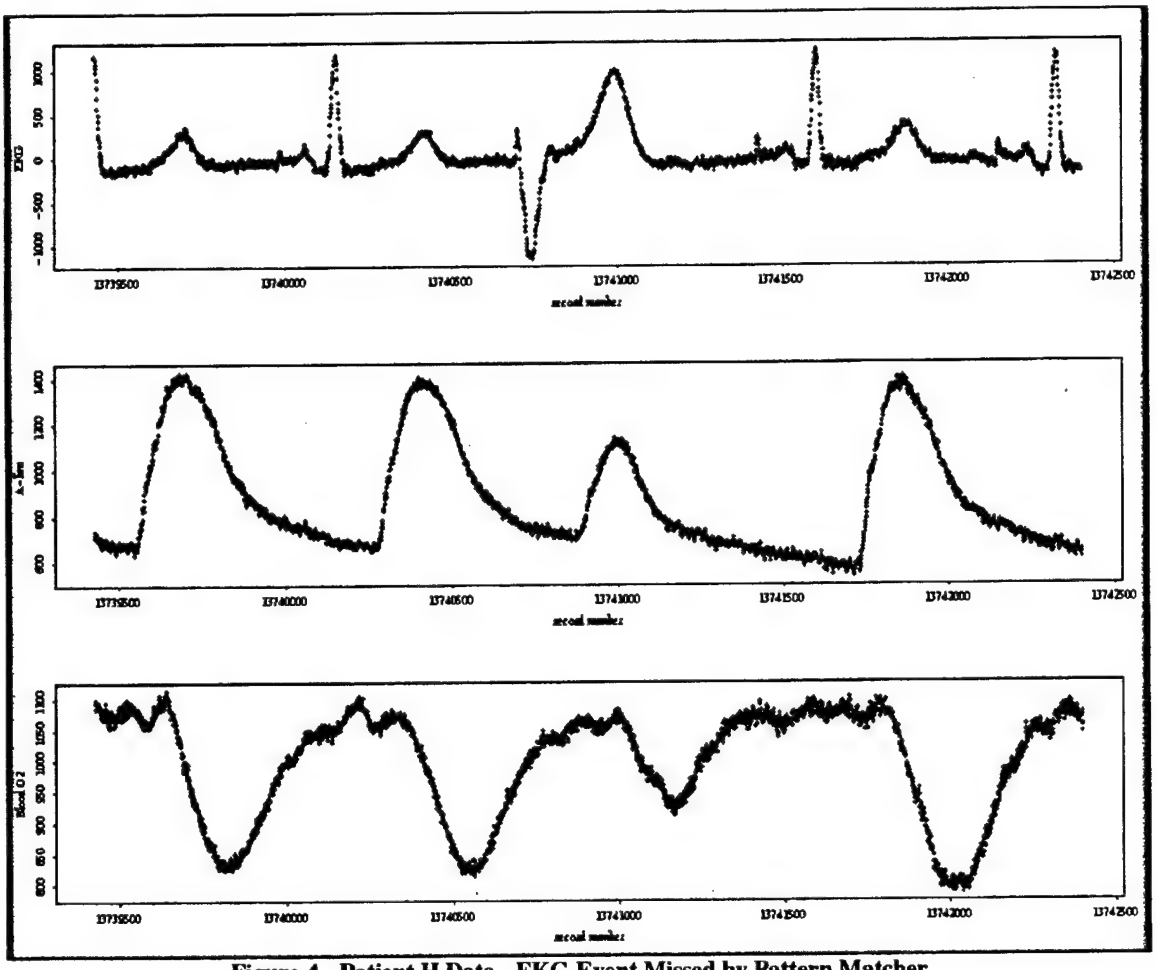


Figure 4 - Patient H Data - EKG Event Missed by Pattern Matcher

EKG, A-line, and Pulse Blood Oximeter data are shown in panels 1, 2, and 3, respectively. The eye can see that something very different has occurred at the third heartbeat. We do not know what caused the inversion of the EKG signal. The heartbeat indicated by the inverted EKG cycle was missed by the pattern matcher, and these types of missing data are NOT shown as a dot in the body of the table. Because the EKG time-pick was missed for this heartbeat, a hole in the summary line at y=0 would occur in the plots contained in Appendix C, "Simple Features – Pattern of Missing Values." For a system such as that described in Appendix G, "MediSense Proposal," it is important for the system to automatically identify time periods where missing data have occurred because of an event so different that the pattern matcher failed to detect it.

Figure 5 shows a span of data from Patient B where the pattern matcher failed to pick out the signature for the third heartbeat in the A-line and Pulse Blood Oximeter data. These are the types of data that are reflected in the figures found in Appendix C.

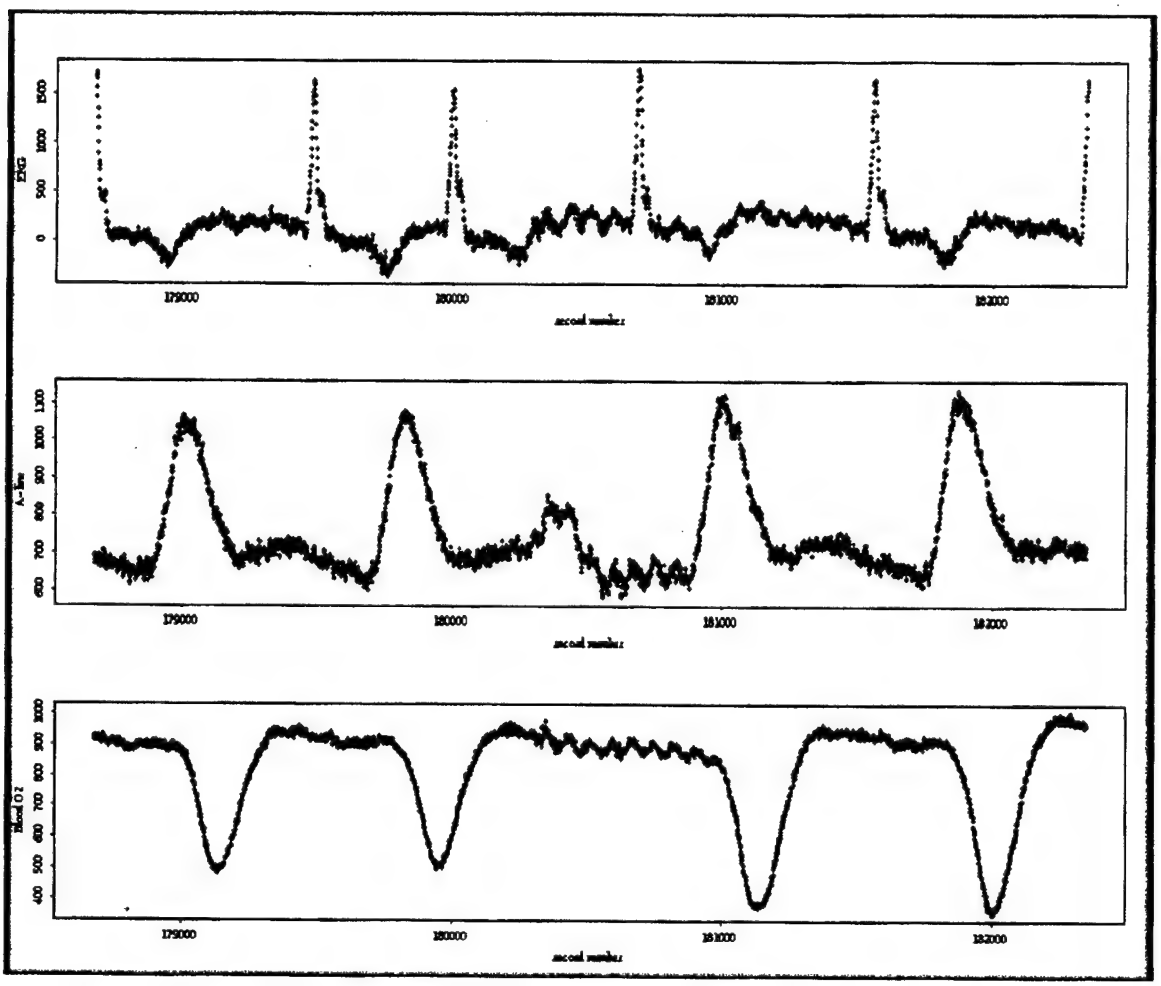


Figure 5 - Patient B Data - A-line and Blood Pulse Oximeter Events Missed by the Pattern Matcher

The strongest patterns that emerge from the pattern of missing values across all patients are the following.

- Inclusion of the "integral" and "donut" features in the analysis cause many events to be omitted from the analysis.
- Patient H has the most missing events.
- Data are omitted through the entire three- to four-hour recording period.

In the MediSense system, it is important to allow the user to select different subsets of variables to be used in the exploratory analysis. For example, an analysis might be performed that includes the "integral" and "donut" features, and another might be performed that excludes these features.

3.4 Cross-Heartbeat Features

The cross-heartbeat features were calculated from statistical summaries of the simple features. Cross-heartbeat features were calculated for each of the simple heartbeat features (variable numbers 4:51 in Appendix B, "Definition of Variable Names") by centering a 60-minute window about each feature time (EKG time, a-line time, and Pulse Blood Oximeter time). If *any* data were available during that time, then three new cross-heartbeat features were calculated: the mean, the standard deviation, and the skewness. (The skewness is a measure of asymmetry of the distribution and is the third moment about the mean.)

3.4.1 Sample Calculations

Figure 6 shows the cross-heartbeat calculations for the A-line Amplitude feature for Patient B. The first three panels show the variable cross-A-line amplitude for mean, standard deviation, and skewness values, respectively. The last two panels display quantities that are in some sense inverses of each other and are diagnostic only. The number of observation averages and the number of missing A-line Amplitude features when a heartbeat was detected in the EKG data (shown in the fourth and fifth panels) are not used in the PCA Analysis found in Section 4.0, "Visualizations of Analysis Results." A break in the signal indicates the absence of detected heartbeats from the EKG data. The fourth panel, after some initial time lag — say, 1000 seconds — is a rough indicator of pulse regularity convolved with missing features.

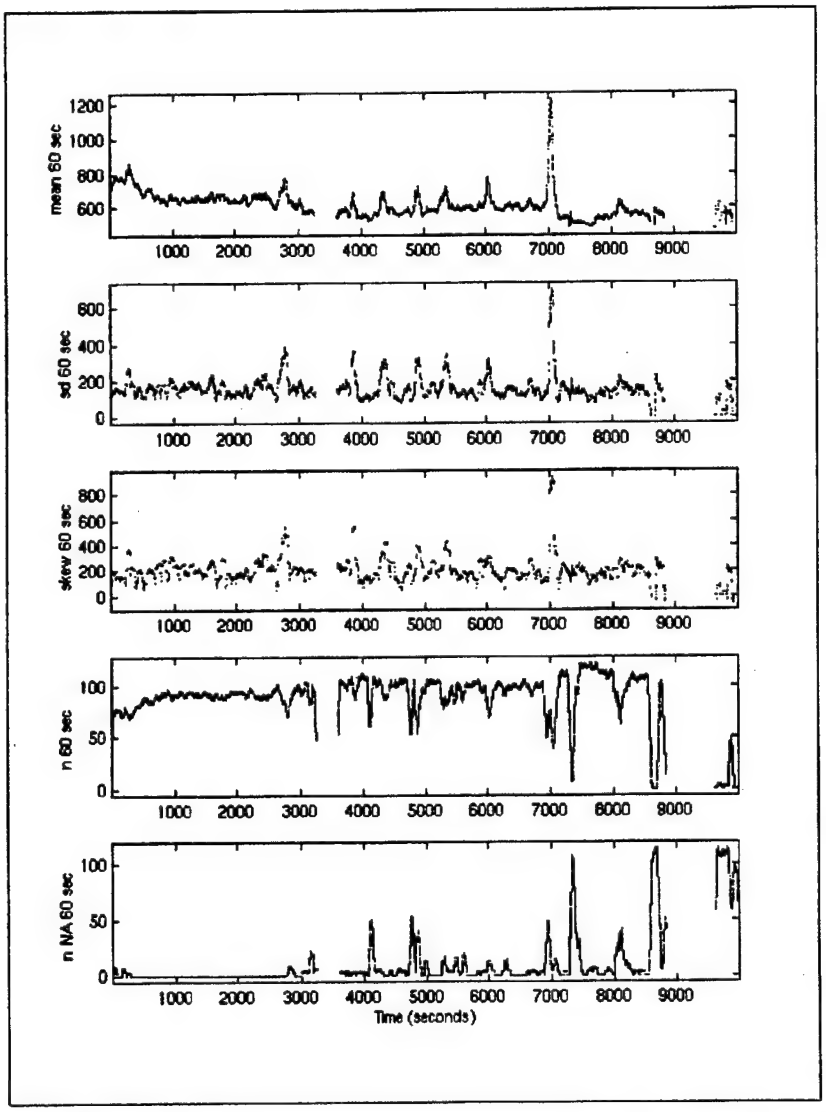


Figure 6 - Patient B Data - Sample Calculations Performed for A-line Amplitude Cross-Variables

Appendix D, "Cross Features – More Sample Calculations," shows a set of these types of plots for each patient for the variable cross-peak R. The fact that Patient H has a pacemaker is very evident from the fourth diagnostic plot when compared to the plots for the other two patients.

3.4.2 Pattern of Missing Values

A graphical summary of all missing data was performed for each of the three patients for the cross-heartbeat features. These plots may be found in Appendix E, "Cross Features – Pattern of Missing Values." The variable numbering used in the plots follows the numbering scheme specified in Appendix B, "Definition of Variable Names." The figures include both the simple and cross-heartbeat features for comparison. The contrast between missing data in the simple and cross heartbeat features is very dramatic because of the way in which we calculate the cross features. Recall that in order for the cross feature to be calculated, it only takes the presence of a single feature someplace in the 60-minute window surrounding a given time.

3.5 Fourier Analysis

Sympathetic and parasympathetic nervous activities are associated with the beat-to-beat heart rate. This association appears in the beat-to-beat variation of the EKG, A-Line, and Pulse Blood Oximeter waveforms. This association may be quantified and tracked in the time domain using features such as the time between peak EKG amplitudes or in the frequency domain using spectral analysis techniques like Windowed Fourier Analysis (WFA). Research indicates that temporal variation in sympathetic and parasympathetic nervous activities can be monitored by tracking the relative power in three power spectrum frequency bands: .01 to .08 Hz (12.5 to 100 seconds), .08 to .15 Hz (6.6 to 12.5 seconds), and .15 to .5 Hz (2 to 6.6 seconds) — in particular, by tracking over time the ratio of low-frequency power to high-frequency power. Computing and tracking this ratio can be accomplished by computing the Fast Fourier Transform (FFT) and then calculating the ratio for a 1-2 minute moving window incrementally advanced over the EKG, A-line, or Pulse Blood Oximeter time series. These computations require significant computing horsepower and computing time to perform directly on time series sampled at 1000 Hz. We did not pursue these calculations because of budgetary constraints. Thus, we concentrated our analyses on the time domain.

3.6 Exploratory Data Analysis

For each patient, PCA was performed using three different but related sets of variables. The three collections of variables were

- Simple - 48 simple heartbeat features (variables 4-51 in Appendix B, "Definition of Variable Names")
- Cross - 144 cross heartbeat features (variables 52-195 in Appendix B)
- Combo - 192 heartbeat features (variables 4-195 in Appendix B).

Each analysis was restricted to a particular patient — no cross-patient analysis was performed. The PCA was performed to reduce the number of variables down to a smaller number for visualization in Data Desk (see Section 4, "Visualizations of Analysis Results"). The theory behind PCA and the loading for the various analyses by patients are discussed below.

3.6.1 Theory - Principal Components Analysis

The objective of PCA is to reduce the dimensionality of a data set that consists of a large number of interrelated variables or features while retaining as much of the variation present in the data set as is possible. This is achieved by transforming to a new set of uncorrelated variables, the Principal Components (PCs), ordered so that the first few PCs explain most of the variation present in all of the original variables. Each PC in the new set is a linear combination of the original variables.

Geometrically, each event in a multivariate data set defines a point in an N-dimensional space whose axes are identified with the N original features. The collection of events, in our case heartbeats, comprises a cloud of points in N-space. Graphically, the first PC axis is the line through the cloud upon which the projected points, the first PC scores, would have the greatest scatter. The second PC axis is the line through the clouds that is orthogonal to the first PC axis and upon which the projected points, the second PC scores, would show the second greatest scatter. This continues on through until the Nth PC axis is found. The origin of the PC axes is located at the centroid of the point clouds.

Mathematically, PCA is accomplished by finding the eigenvectors and eigenvalues of the covariance matrix of the original N variables if the scales are directly comparable. We used the correlation matrix because the scale of the

variables was not directly comparable. Each eigenvector provides the coefficients of the linear combination of the original variables that defines one line (PC axis) through the clouds. The eigenvalues are the variances of the projected points (PC scores) onto the lines (PC axes).

3.6.2 Loadings by Analysis Type

In this section, we examine trends in the variables that have the largest loadings and contrast the percent variation that has been explained by the three variable collections for each patient and for the first few PCs. More detailed information is also available in Appendix F, "PCA Loadings."

3.6.2.1 Variable Loading on PCs

In Appendix F, "PCA Loadings," tables are provided for each patient by variable collection. These tables specify the 10 variables with the largest absolute value for their loading coefficients for the first six PCs. Some general trends emerge from these tables.

For the simple variable collection, the first PC tends to be dominated (has coefficient with the largest absolute value) by the A-line variables. This is true across patients. The actual variables with the largest loading vary by patient, but there are strong similarities across patients, especially between Patients B and W. The variable with the largest coefficient for Patients B and W is "high d1 Aline". For Patient H, this variable is the fifth most important. The "integral Aline" is the strongest for Patient H. The second PC is dominated by the Pulse Blood Oximeter variables. The variables with the largest loadings again show strong similarities across patients, but Patients H and B appear more similar in this PCA than does Patient W. The third PC tends to be dominated by Pulse Blood Oximeter variables or a combination of Pulse Blood Oximeter and A-line features as is the case for Patient B.

For the cross-feature variable collection, the trends observed for the simple variables do not hold. The coefficients from the spline fits for the A-line and Pulse Blood Oximeter "donut" are much more prominent. For Patient H, the key variables for the first two PCs are uniformly from the spline fits with the variable from the first component primarily derived from the Pulse Blood Oximeter variable. For Patients W and B, we do not see "donut" coefficients dominating the first two PC calculations. For Patient W, the Pulse Blood Oximeter variables and A-line variables tend to dominate the first and second PCs, respectively. This is the opposite of what we saw in the simple analysis. For Patient B, we see the same pattern of variable composition: A-line and Pulse Blood Oximeter variables for the first two PCs.

For the combined variable analysis for Patient H, we see a combination of the behaviors noted above: PC1 is dominated by A-line variables and PC2 is dominated by Pulse Blood Oximeter variables — but now there is a balance of feature geometry and "donut" spline coefficients. We see an analogous pattern for Patients B and W.

Appendix F, "PCA Loadings," contains graphs that show the magnitude and the sign of the loading coefficients for the first six PCs for each of the patient/variable combinations.

3.6.2.2 Variation Explained

Tables 2 - 4 below summarize the percentage variance explained by the first six PCs for the variable collections used in the three analyses: simple, cross, and combined by patient. In these tables, the column "EigVal" is the eigenvalue associated with the PC; this is the " PC Variance" because the eigenvalue is the variance of the associated PC scores. The second column, entitled "%Var," is the percentage variance explained in the original variables, and the last column, "Cum %Var," simply accumulates these percentages.

For the simple variable collection shown in Table 2, the first three PCs explain about half of the variation. The pacemaker patient, H, requires four PCs to explain roughly the same variation as is explained by the first three PCs for the other two patients, possibly indicating an additional confounding factor not taken into account by the variables available to the analysis.

Table 2 - Simple PCA Loadings

PC	*EigVal*	*%Var*	*Cum %Var*
	Patient	**Patient**	**Patient**
	B H W	**B H W**	**B H W**
1	12 11 15	24 24 32	24 24 32
2	9 7 9	18 14 18	42 37 49
3	6 5 6	12 11 13	54 48 63
4	5 4 3	9 9 6	64 58 68
5	3 3 2	7 6 4	70 64 72
6	3 2 2	6 5 3	76 68 75

The cross-variable loadings are shown in Table 3. Interestingly, in the cross-variable analysis, the loadings for the pacemaker patient, H, are not distinguished from the other two patients. One could hypothesize that the irregularities in heartbeat that have influenced the calculation of features in the simple variable collection are damped in the 60-minute averaging windows that have been applied to create this new cross-variable set. As a general observation, the amount of variation explained by the first five PCs in the cross analysis is roughly comparable to that which is explained by the first three PCs in the simple analysis. This seems to indicate that there is some justification for combining the simple and cross variables into another set for analysis.

Table 3 - Cross PCA Loadings

PC	*EigVal*	*%Var*	*Cum %Var*
	Patient	**Patient**	**Patient**
	B H W	**B H W**	**B H W**
1	33 31 26	23 22 18	23 22 18
2	17 17 20	12 12 14	35 33 32
3	12 13 13	8 9 9	43 42 41
4	10 10 10	7 7 7	50 49 48
5	8 9 9	6 6 6	56 55 54
6	7 8 6	5 6 4	61 61 59

The loadings for the combined variable set are shown in Table 4. In the combined variable analysis, the pacemaker patient, H, again appears a bit more complicated that the other two patients. Four PCs are required to explain 50% of the variation for Patients B and W, while five are needed to explain 50% for Patient H.

Table 4 - Combo PCA Loadings

PC	*EigVal*	*%Var*	*Cum %Var*
	Patient	Patient	Patient
	B H W	B H W	B H W
1	49 37 40	26 19 21	26 19 21
2	24 21 27	12 11 14	38 30 35
3	14 17 17	7 9 9	45 39 44
4	11 12 12	6 6 6	51 46 50
5	8 10 12	4 5 6	55 50 56
6	7 8 8	4 4 4	59 55 61

In all cases — simple, cross, and combined — the incremental contributions in explained variance is greatest for the first three PCs.

4.0 Visualizations of Analysis Results

The PCA scores for each captured event for the first six PCs for the simple, cross, and combined analyses were loaded into Data Desk for each patient. Two other quantities, which were not used in the PCA, "EKG time" and "Peak R N," are used to color-code the plotted PCA event scores according to temporal position. "EKG time" is the number of seconds into the data record at which time the peak R value was detected. Events early in the data record are colored with cool (blue) colors; later events are colored hot (red). This same color-coding is used throughout this report.

4.1 Heart Rates

The choice of "Peak R N" is somewhat arbitrary but does contain some useful information. "Peak R N" is the number of observations that were available for the moving window calculation for the cross variables:

67 peak R mean 60 sec
68 peak R sd 60 sec
69 peak R skew 60 sec

Figures 7 - 9 show the plot of "Peak R N" versus "EKG time" for each of the three patients. The ordinate values are affected by both the patient's heart rate (beats per minute) and amount of missing data and/or data for which the EKG Peak R feature was not detected. A reference back to Figure 6 in Section 3.3.2, "Multiple Sensor Features," clearly shows the correspondence between missing values and big excursions (dips) from a heart-rate line.

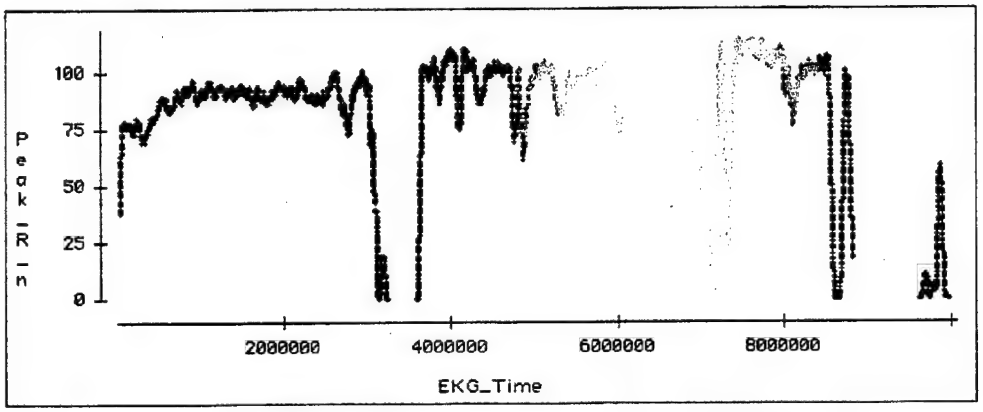


Figure 7 - Patient B Data - Plot of "Peak R" vs. "EKG Time"

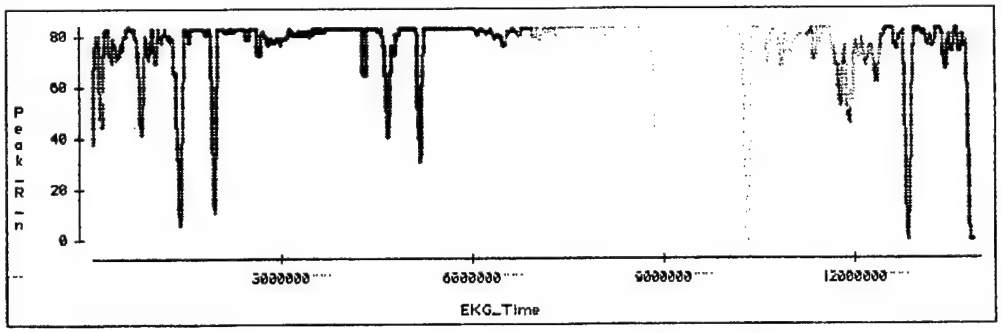


Figure 8 - Patient H Data - Plot of "Peak R" vs. "EKG Time"

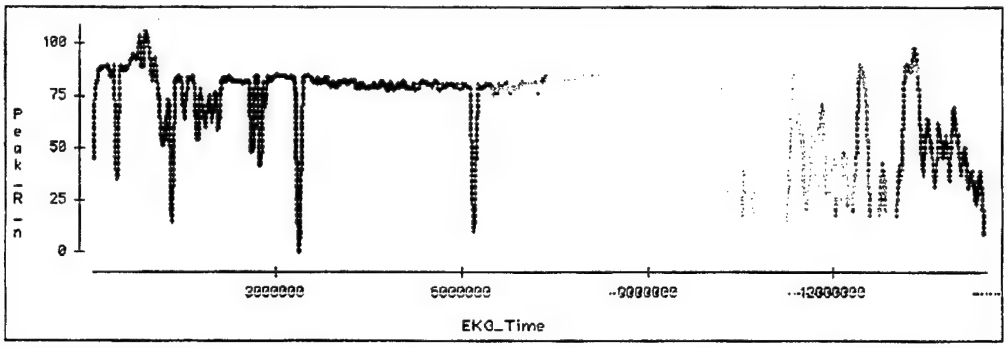


Figure 9 - Patient W Data - Plot of "Peak R" vs. "EKG Time"

The large spikes downward in amplitude for the "Peak R N" values for Patients H and W correspond exactly to the chunks of missing features as can be seen in Appendix D, "Cross Features – More Sample Calculations." The smaller amplitude wiggles are more likely due to variations in heart rate. The stability of the heart rate for the pacemaker patient shown in Figure 8 is readily apparent. Accounting for missing data, Patient B seems to have the greatest variation in heart rate — a much lower rate initially and then gradually increasing.

All patients shown an initial spike downward because of the way in which the averages were defined.

4.2 PCs and Time

Figure 10 below shows the first three PCs for Patient H plotted against "EKG time" for the combined variable set analysis. The color-coding remains the same as in all the figures used in this report. Note that a PC is not the same as time and that there is an interesting relationship among the first three PCs over time for this patient. The relationship will of course be different for each patient and variable set.

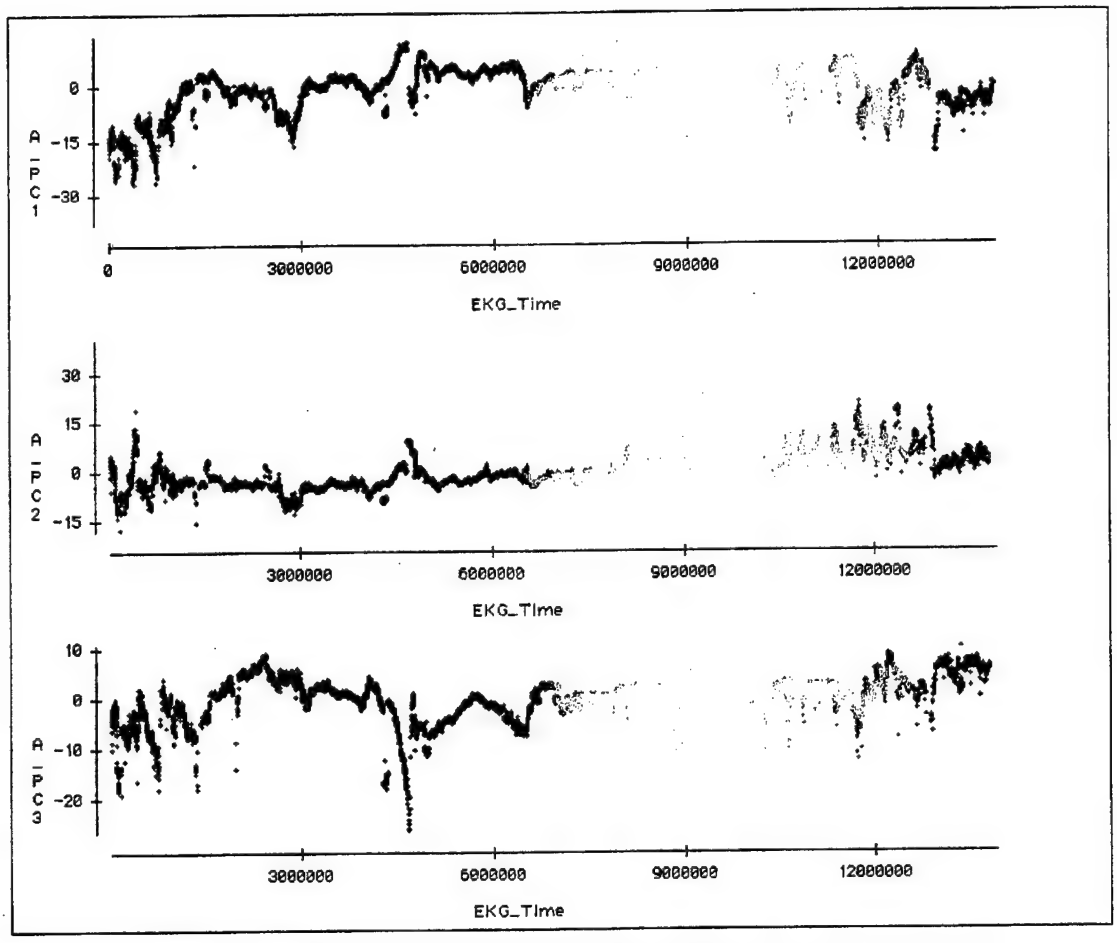


Figure 10 - Patient H - First Three Principal Components Plotted Against "EKG time"

4.3 Some Pairwise Comparisons

All pairwise combinations for the first four PCs from the combined variable set are show in Figure 11 for Patient H. The color time-coding helps to identify the very different "behavior" that the heartbeat events have under the various PC projections. For example, observe how the green events appear as outliers in the "PC1 vs PC2" and "PC1 vs PC3" plots but are tightly bound together in the "PC1 vs PC2" plot. In an interactive system such as that described in Appendix G, "MediSense Proposal," the user of the system could explore the original data streams for those times, the features derived for the highly "loaded" variable, and any ancillary data available for the patient. The balance of the pairwise plots for the remaining eight patient/variable combinations is not shown in this report. All are different; all are intriguing. The next section helps quantify how very different the visualizations are for the various patients.

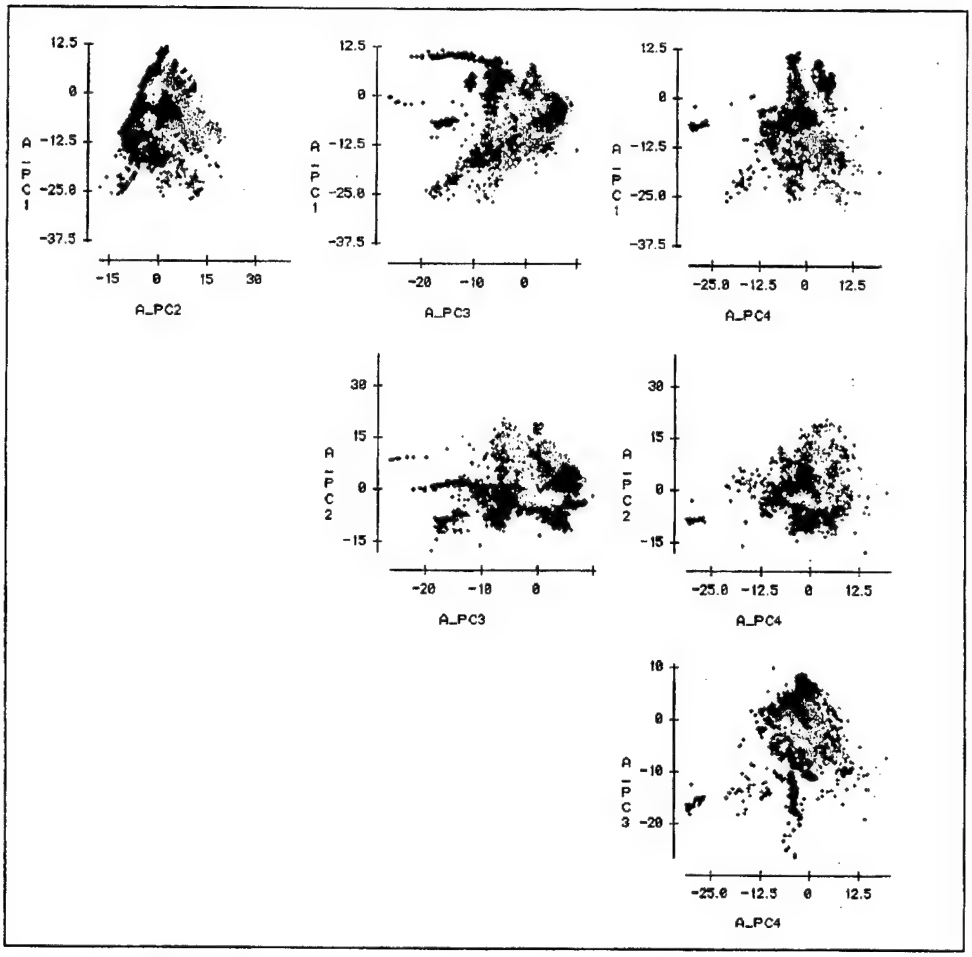


Figure 11 - Patient H Data - Pairwise Plots for the Combined Variable Set Using the First Four PCs.

4.4 Comparisons Across Variable Collections by Patient

The very best way to view the PC projection plots is to use a 3-D viewer with animation and interaction. Included with this report is a VCR tape that shows several of the analyses using Data Desk's rotating plot graphics. Only a few of the possible views of the PC scored data are presented for the three patients below in Figures 12 - 14. The pictures are meant to be tantalizing rather than comprehensive.

The format is the same in each of the subsequent figures. The first two PCs are shown for each of the patients for the simple, cross, and combination analysis. The simple variable projections are the top panels; the cross-variable projections are the middle panels; and the combined variable projections are on the bottom.

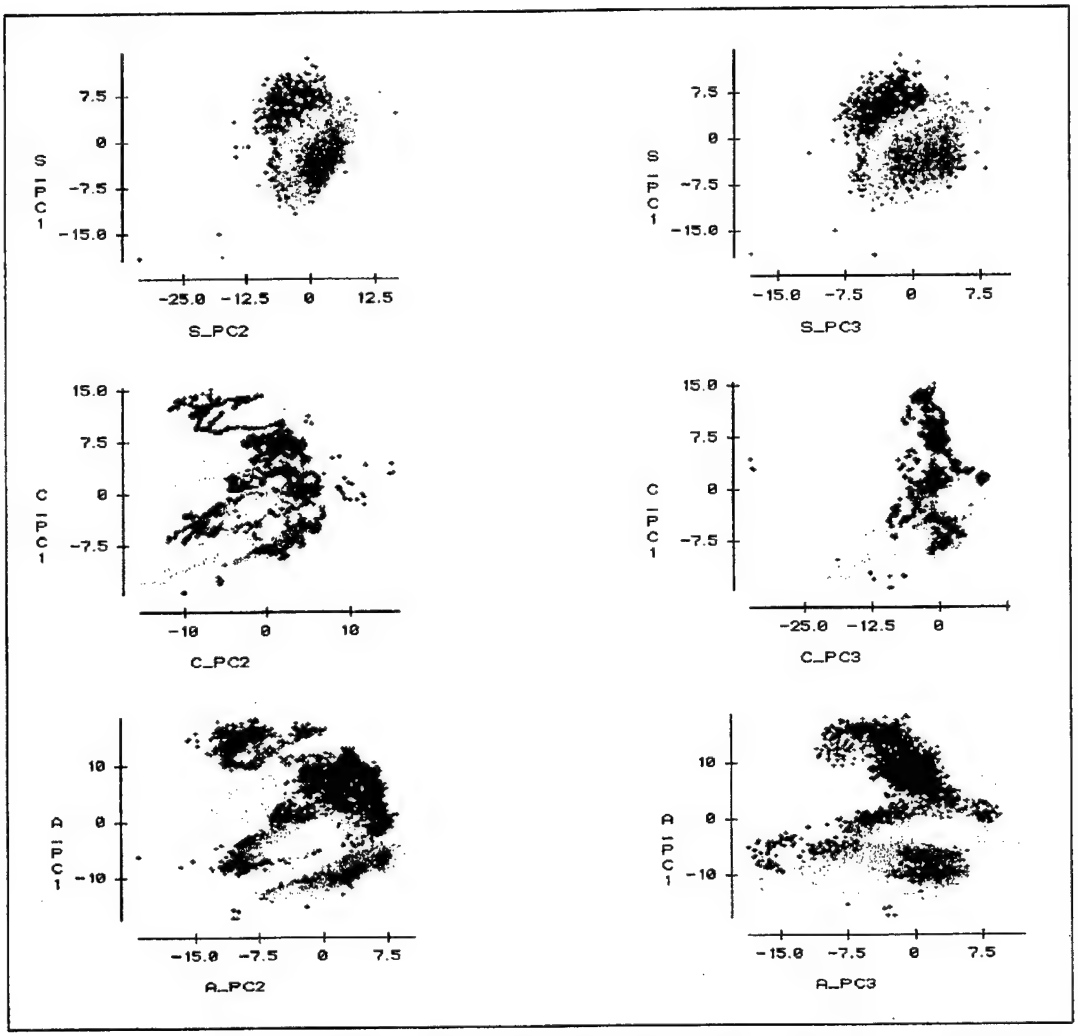


Figure 12 - Patient B Data - Simple, Cross, and Combined PC Projection Plots

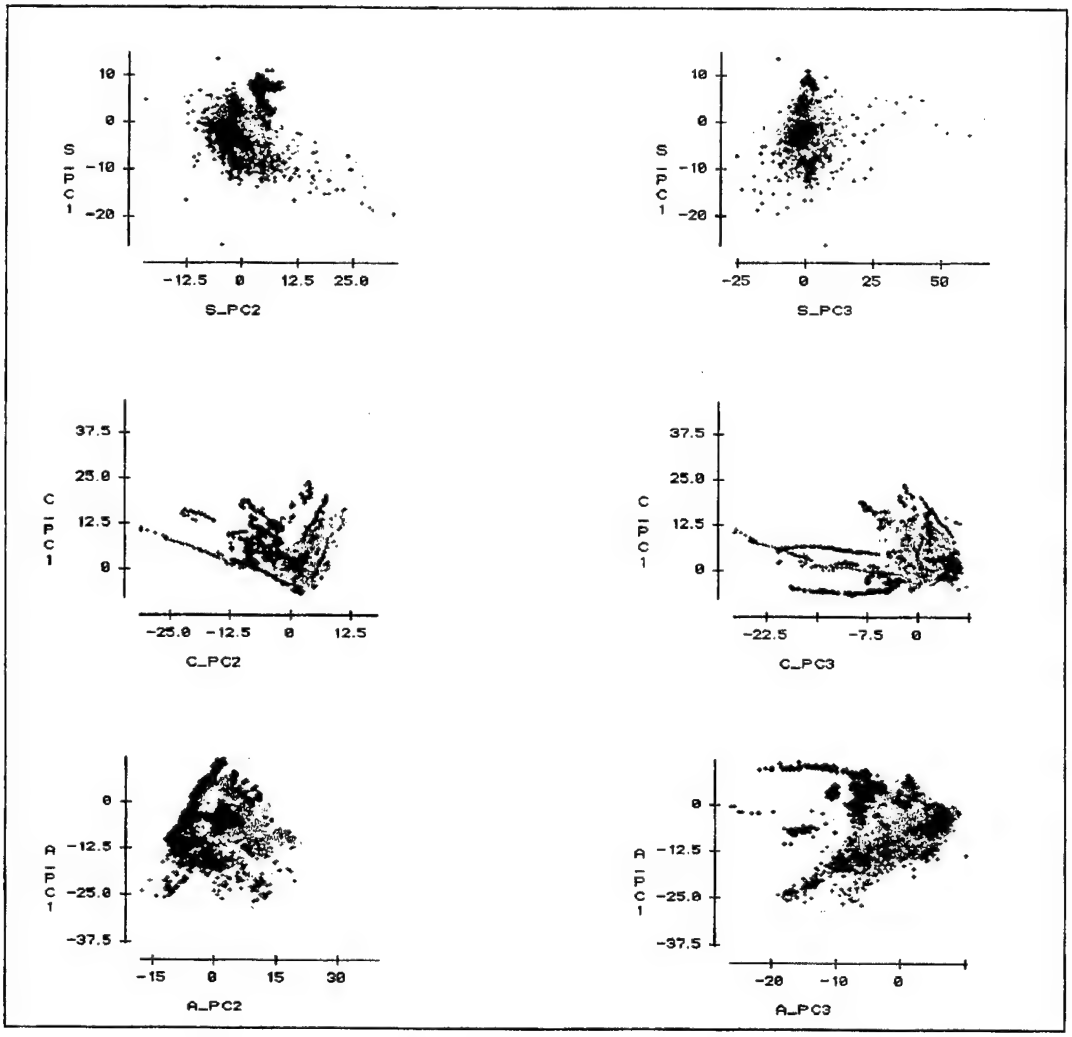


Figure 13 - Patient H Data - Simple, Cross, and Combined PC Projection Plots

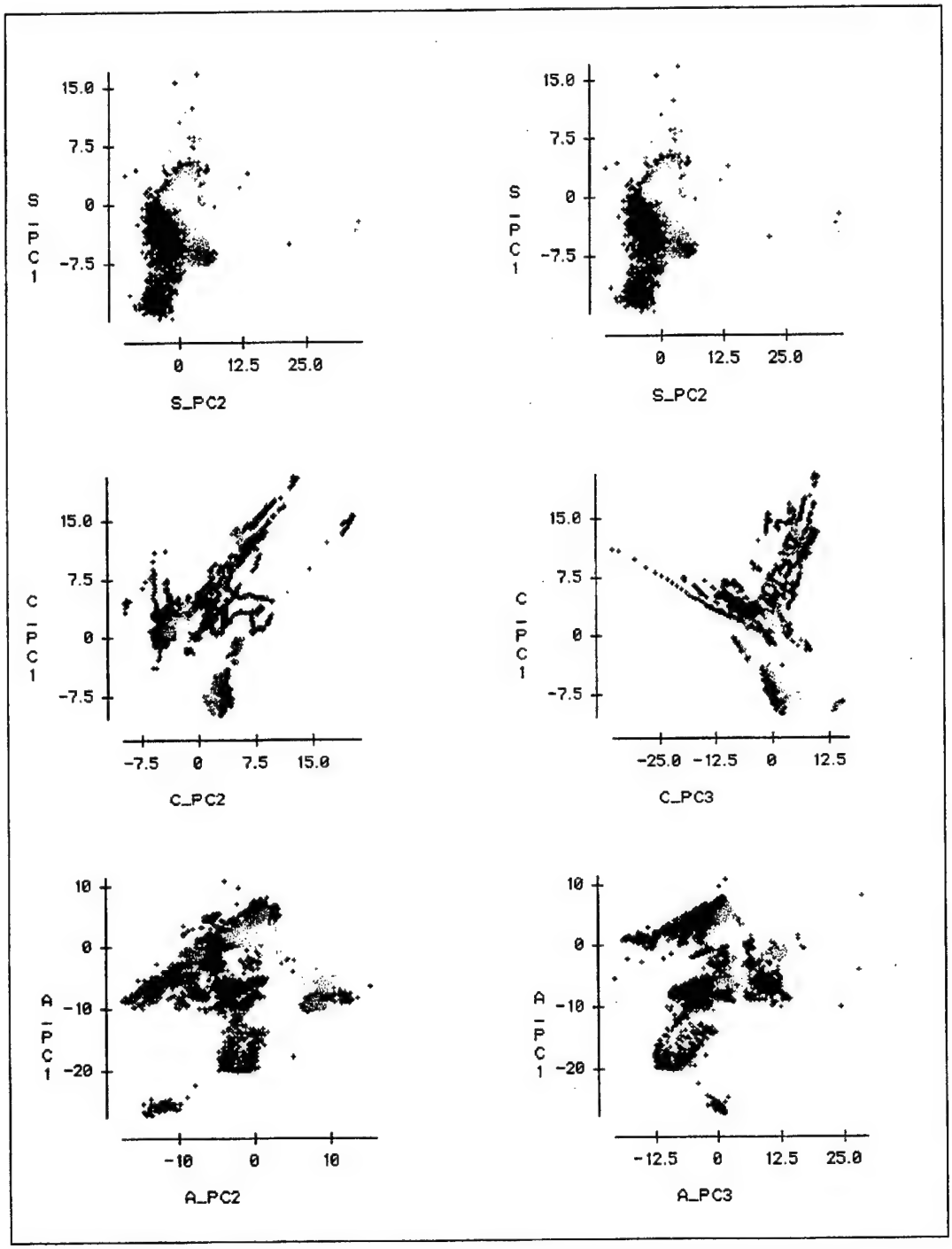


Figure 14 - Patient W Data - Simple, Cross, and Combined PC Projection Plots

5.0 Proposed Future Work

Two types of future work are proposed. The first is a small task to complete a frequency domain analysis for the current data set using WFA or Wavelet Analysis (WA). The second, larger piece of follow-on work is to actually build an interactive visualization system where the user can investigate interesting branches of the data at fine levels of detail and drive the exploratory data analysis by selecting which collection of variables to include in the PCA, for

example. Although ancillary data were not available in our study to correlate to the sensor data, this would be very easy to do in an interactive visualization system.

5.1 Frequency Domain Analysis

Classical spectral analysis is impractical for huge data sets. Other, more practical approaches can provide similar information and more. These approaches include WA and Dynamic Linear Modeling (DLM). The first, WA, is better suited to the analysis of time series featuring transient behaviors. WA came into being partly to overcome the shortcomings of WFA. In one sense, it is a natural generalization of Fourier Analysis and WFA for the analysis of time series data whose spectral content varies with time. DLM is a time domain method whose computation is very efficient and may be implemented in real time because it makes use of each observation as it is sampled. It can also be configured to closely monitor spectral variation over time.

5.2 MediSense

The proposed MediSense system will provide near-real time data analysis of multi-sensor medical data for a given patient and across patients. The system would be implemented in C++ and would be optimized to run quickly. The full proposal is in Appendix G.

5.2.1 Background

Multi-sensor medical data from patients are increasingly available. An example is the Life Support for Trauma and Transport (LSTAT) system that will house a variety of cardiac and respiratory sensors. The data from these sensors are potentially useful for real-time monitoring of patient conditions, diagnosing chronic health problems, and triage. However, because of the size and complexity of the data, it is difficult to survey even a few hours of sensor data, much less use the data for the three items above. Based on last fiscal year's work, we think that this project will take a giant step forward in examining these types of data.

This work is a continuation and enhancement of the Signal Processing Concept Exploration Task that is part of the Information Analysis and Visualization Research and Development Project for the U.S. Army and Material Command. Our clients are Major Stephen Bruttig of MRMC and Dr. Frederick Pearce of WRAIR Surgery Division. Based on work in fiscal year 1997 (FY97) and the large amounts of data that the LSTAT project will generate, it may be more appropriate for this task to become its own project. The system we propose to develop is called ***MediSense***. Aspects of the envisioned system are "SPIRE-like" in that the data analysis is unsupervised and high-level features can be automatically extracted by the system, but the SPIRE system is not used.

5.2.2 Goals

The long-term goals of this work are to combine the data from a variety of real-time medical sensors from a single patient or from multiple patients so that

- a patient's detailed condition can be readily ascertained by medical care providers
- a patient's sensor history can be quickly reviewed by the medical care providers
- multiple patients' physical states can be assessed and compared, to prioritize and group the patients for efficient treatment of injuries.

The targeted application is the multiple sensors of the LSTAT stretcher.

5.2.3 Technical Approach

To achieve the above goals, this year's work will focus on the following:

1) **Detailed analysis and presentation of a patient's condition.**
   This task examines LSTAT-like data from several patients at the heartbeat-to-heartbeat level of detail, over a specified time period of several hours or more. This work is a continuation of work begun in FY97. Activities that are included in this work are

a) Calculating summaries and "low-order" features from the individual sensors. For instance, this year we developed algorithms to detect individual heartbeats in three sensors and then calculated summaries related to the "shape," spacing, and location of the heartbeat features.
b) Detecting events in individual sensor data streams. For instance, this year we detected the absence of a clean heartbeat feature when such a feature was expected.
c) Reducing dimension/data. Ways to view a few hours of a patient's sensor data, based on the summaries and events described in parts a) and b), will be explored. Alternatives will be evaluated and presented. A possibility for the data reduction is organizing for each patient the collection of features from the multiple sensors into an event feature matrix where time is one dimension and derived features comprise another dimension. A statistical analysis, such as PCA, is used to reduce the dimensionality of the event space to two or three. This allows the events, which are chunks of time from the patient sensor records, to be projected into "diagnostic-space."

2) **Visualization of the patient's current state.**
We will create prototypes of displays that can be used to summarize (in real time) a patient's condition. These displays may be a simple icon, glyph, or other graphic. The displays will be based on the features derived in Task 1. Simple interactions (designed to reveal more information as needed by the attending medical personnel) with the displays will also be prototyped.

3) **Cross-patient compressed views.**
This task provides for a cross-patient visualization that spatially organizes the individual icons in "diagnostic or alarm" space. Several views are possible. A geometric-based view places the summary icon for a patient in a location on the screen suggested by where the patient is physically located. A diagnostic-based view places patients whose conditions are most similar close together on the screen. In any case, various conditions noted in the sensor data would signal the icon to go into alarm mode.

4) **Architectural issues for the MediSense system.**
This task plans for the future by providing broad software and hardware architectural guidance for the **MediSense** analysis platform, as it will apply to LSTAT (or related) sensor data. Getting this task underway during this fiscal year will allow us to better leverage our exploratory work in the construction of the eventual product.

5) **Handling dropped sensors or other missing/corrupt data.**
It is inevitable that either a sensor(s) will fail or the sensor suite in LSTAT will evolve. This task will explore how such inevitabilities will be handled; in particular, for many patient states, it may be that useful monitoring or triage decisions can be made with some subset of the full suite of sensors.

6) **Relationships between patient state (e.g., satisfactory, dying etc.) and sensor features.**
The relationships between patient states and features derived from the sensors must be explored. For instance, what sensor states correspond with an individual who can be safely left alone for an hour or two? What sensor states correspond with an individual that is in immediate need of active care? What sensor states correspond with an individual who is beyond medical science's ability to resuscitate? These relations between patient states and sensor measurements must be made for MediSense to be a diagnostic system (however, exploratory analysis of the data in Task 1 does not require such relations to be known). This task will gather together readily available relations and provide a plan for how this necessary relation can be further elucidated. Note that standard options for obtaining such a relation range from using available observations (e.g., we can use existing data to learn what some cardiac data looked like for individuals who lived at least another hour) and designing animal experiments.

7) **Presentation of results in an open forum.**
Systems like LSTAT are creating new data. The combination of the data from the multiple sensors that will be informatively combined in Task 1 is effectively a new medical measurement. New data lead to new science. An important activity will be to present the information created in Task 1 to medical and physiological researchers, so that we may leverage the knowledge in the research community to improve the diagnoses made based on the new data being generated in this project.

This system is discussed in more detail in Appendix G, "MediSense Proposal."

6.0 References

Data Desk Ver. 6.0. Data Description, Inc. Ithica, NY

de Boor, C. "A Practical Guide to Splines." 1978. *Applied Mathematical Sciences*, vol. 27. Springer-Verlag, New York.

Little RJ and DB Rubin. *Statistical Analysis with Missing Data*. 1987. John Wiley & Sons, New York.

MATLAB, Ver. 5.0. The MathWorks, Inc. Natick, MA.

S-PLUS, Ver 4. Statistical Sciences, Inc. Seattle, WA.

Appendix A

Medline Corpus Visualizations

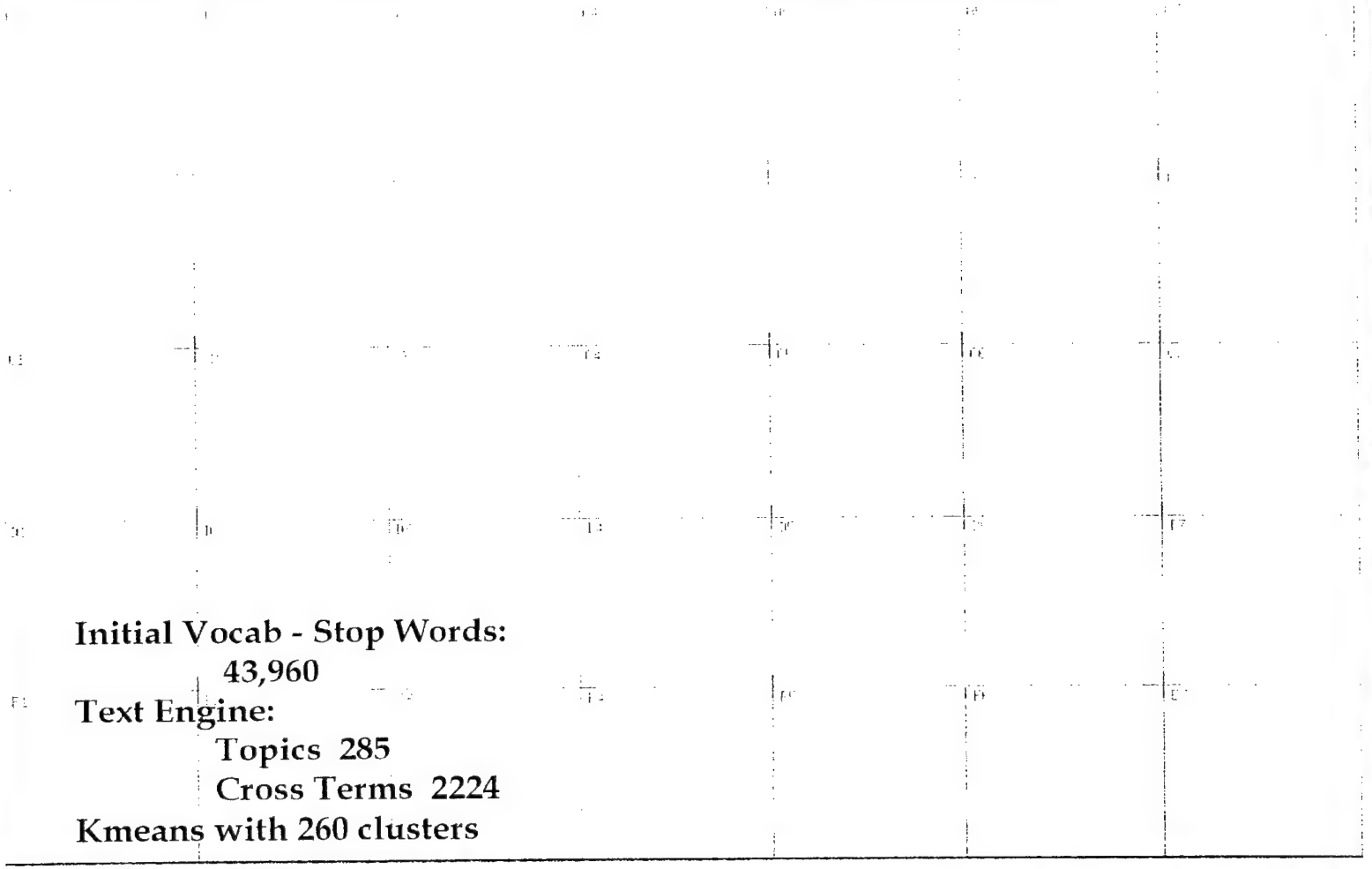
Medline Abstracts
National Library of Medicine

- Publication 1994 to 1997
- Query Hemorrhage or Shock
- Retrieval 18,000 Titles (~1% empty)
- Abstracts Over 16,000 retrieved
- Availability Research Use Only

ARMY Signal Processing Concept Exploration

Medline 16,000+ abstracts

GALAXY View



Initial Vocab - Stop Words:
43,960

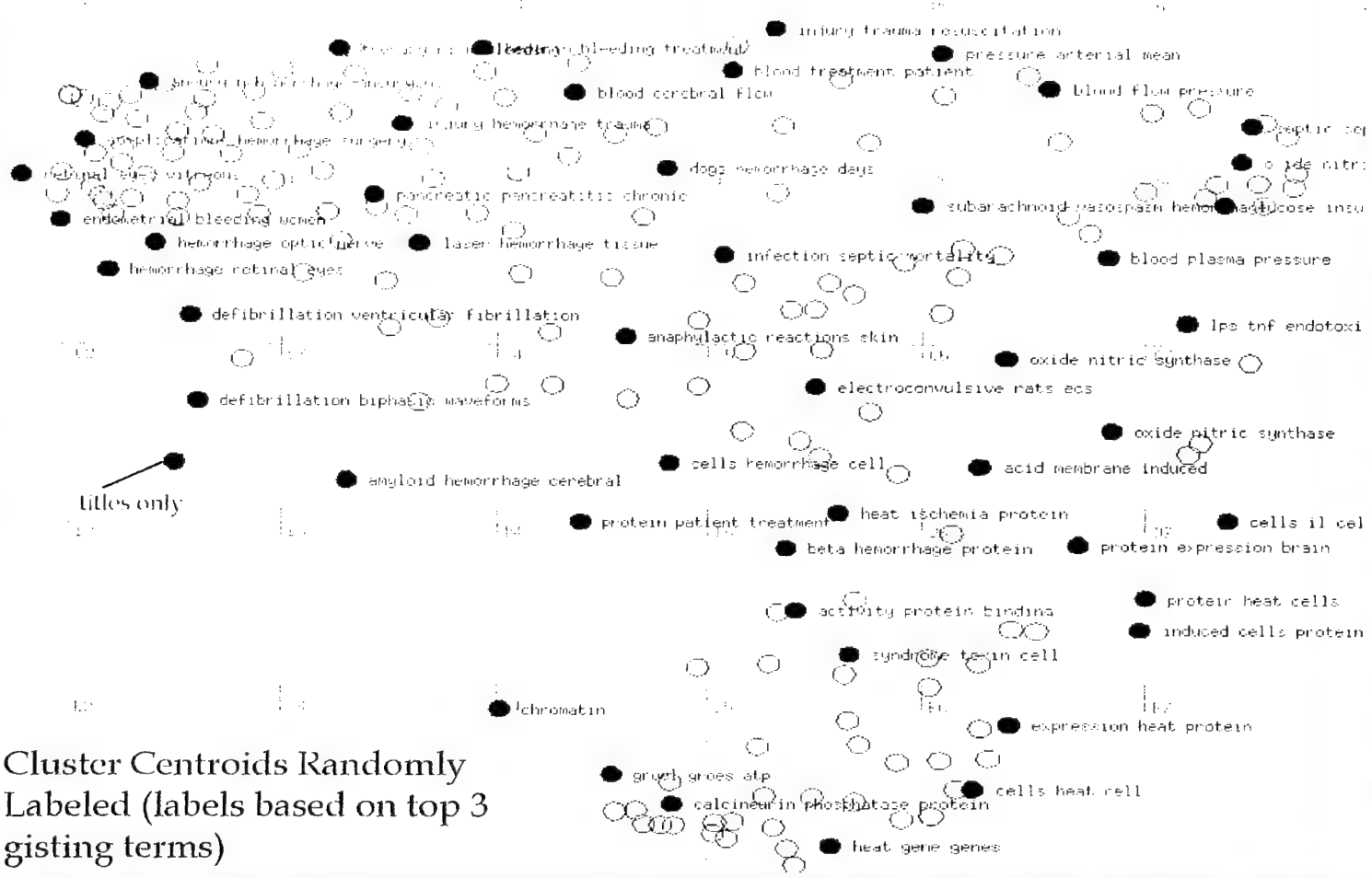
Text Engine:
Topics 285
Cross Terms 2224

Kmeans with 260 clusters

ARMY Signal Processing Concept Exploration

Medline 16,000+

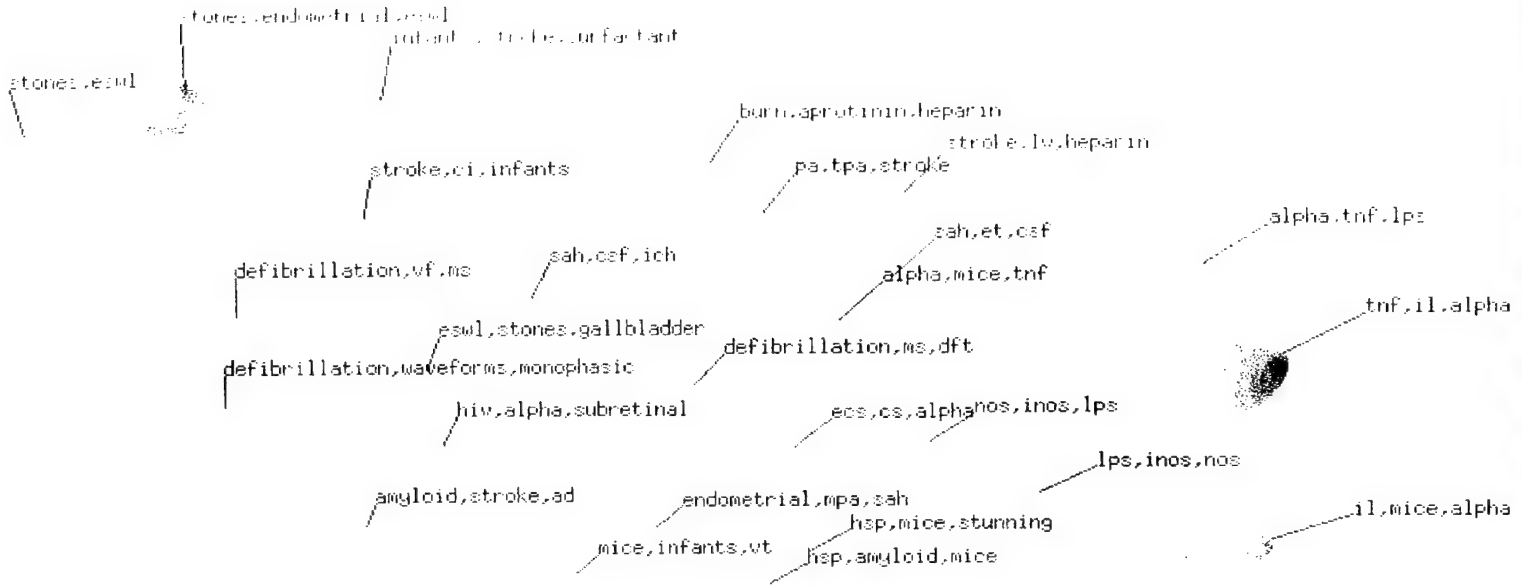
GALAXY Centroids



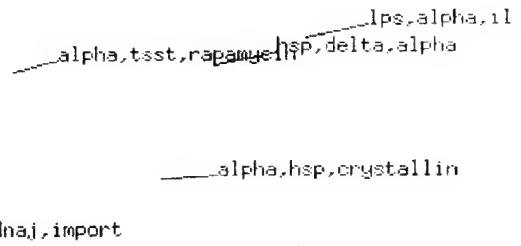
Cluster Centroids Randomly Labeled (labels based on top 3 gisting terms)

Medline 16,000+

Themescape View



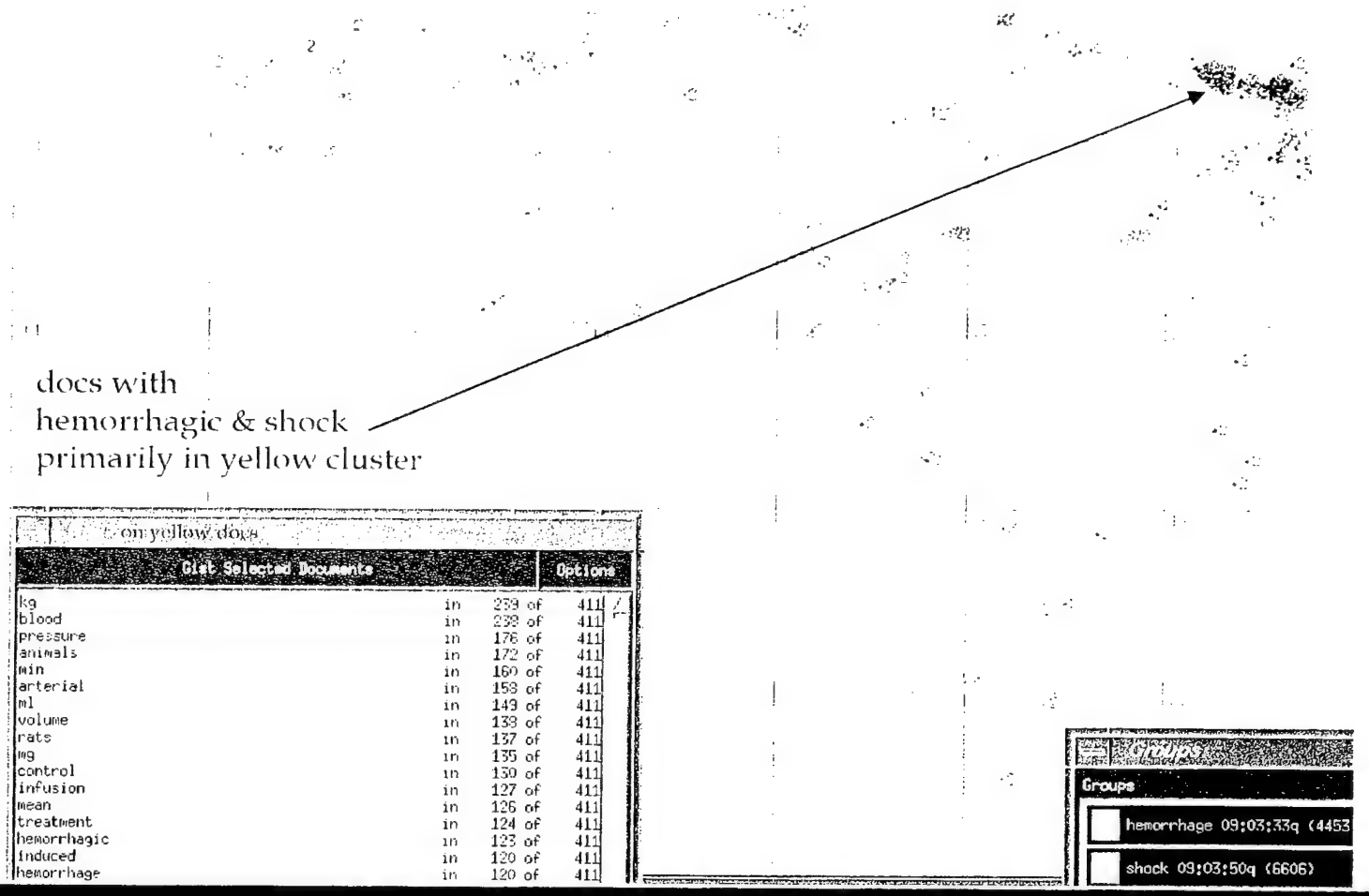
Floor is galaxy plane,
Height is smoothed
cumulative topic count
for top corpus topics



ARMY Signal Processing Concept Exploration

Medline 16,000+

Defined Groups: pink-> hemorrhage blue->shock



ARMY Signal Processing Concept Exploration

Medline 16,000+

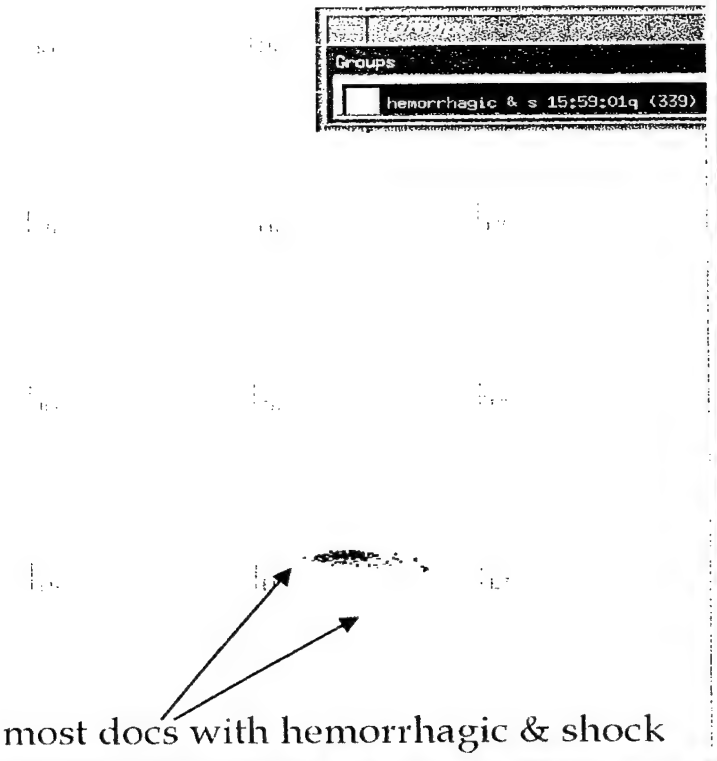
NEW User Weighted View

Gist on yellow upper cluster

Options

blood	in	283 of	369
pressure	in	230 of	369
arterial	in	210 of	369
mean	in	171 of	369
volume	in	154 of	369
resuscitation	in	136 of	369
ml	in	135 of	369
kg	in	134 of	369
cardiac	in	128 of	369
min	in	127 of	369
animals	in	123 of	369
flow	in	120 of	369
hemorrhage	in	116 of	369
hemorrhagic	in	113 of	369
control	in	93 of	369
mm	in	91 of	369
oxygen	in	89 of	369
hg	in	87 of	369
infusion	in	87 of	369
output	in	85 of	369
fluid	in	84 of	369
baseline	in	84 of	369
treatment	in	82 of	369
induced	in	81 of	369
systemic	in	80 of	369
saline	in	79 of	369
values	in	75 of	369
hours	in	72 of	369
rate	in	71 of	369
solution	in	66 of	369
received	in	64 of	369
pulmonary	in	64 of	369
heart	in	63 of	369
injury	in	62 of	369
rats	in	62 of	369
venous	in	62 of	369
artery	in	62 of	369
tissue	in	58 of	369
treated	in	56 of	369
index	in	56 of	369
mg	in	55 of	369
minutes	in	55 of	369
ringer	in	54 of	369
vascular	in	52 of	369
perfusion	in	51 of	369
mmhg	in	50 of	369
resistance	in	49 of	369
levels	in	48 of	369
trauma	in	48 of	369
plasma	in	48 of	369
acute	in	47 of	369
administration	in	46 of	369
lactate	in	44 of	369
hypertonic	in	43 of	369
delivery	in	43 of	369

weighted terms: resuscitation, kg, plasma, oxygen



ARMY Signal Processing Concept Exploration

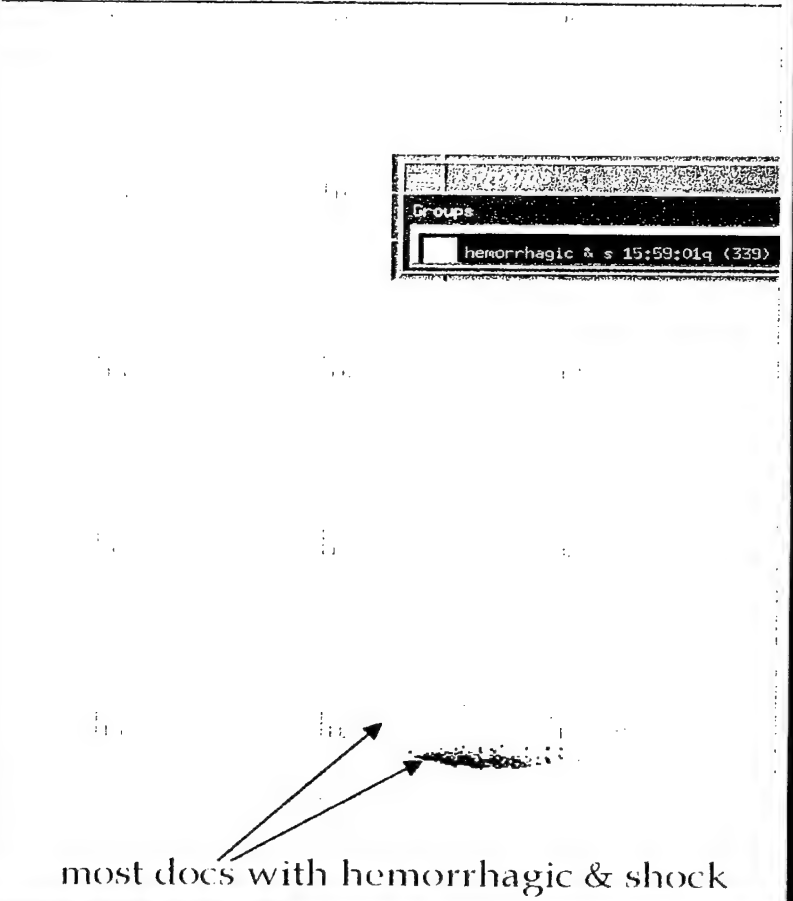
Medline 16,000+

NEW User Weighted View

Gist on yellow lower cluster — Options

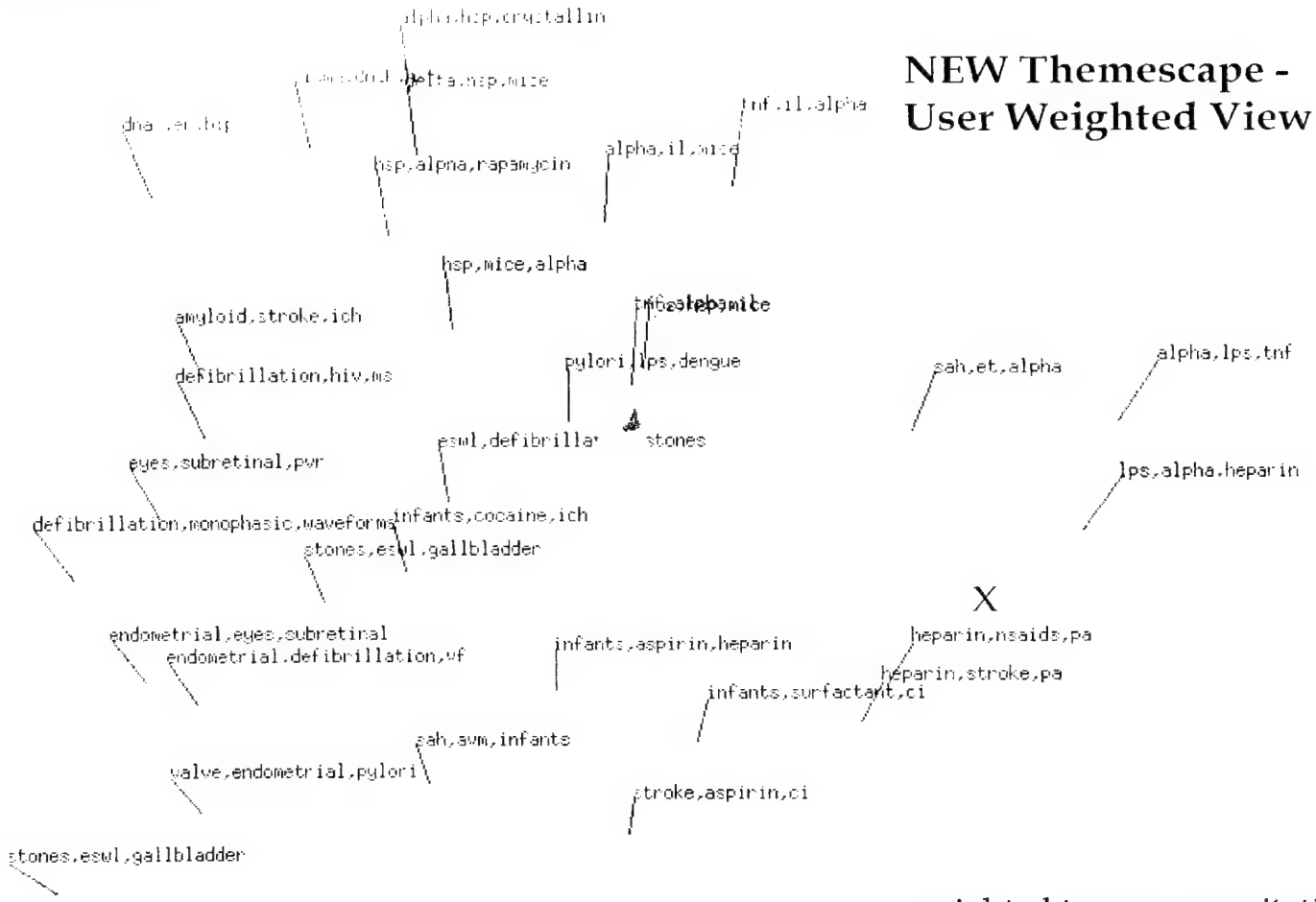
weighted terms: resuscitation, kg, plasma, oxygen

blood	in	292 of	490
pressure	in	214 of	490
arterial	in	179 of	490
mean	in	165 of	490
min	in	159 of	490
kg	in	156 of	490
control	in	147 of	490
ml	in	145 of	490
animals	in	139 of	490
hemorrhage	in	136 of	490
induced	in	135 of	490
rats	in	128 of	490
treatment	in	126 of	490
mg	in	116 of	490
levels	in	113 of	490
administration	in	105 of	490
treated	in	103 of	490
infusion	in	100 of	490
hemorrhagic	in	95 of	490
cardiac	in	88 of	490
rate	in	87 of	490
flow	in	87 of	490
volume	in	86 of	490
injury	in	86 of	490
plasma	in	84 of	490
bleeding	in	81 of	490
hours	in	81 of	490
artery	in	78 of	490
systemic	in	77 of	490
received	in	74 of	490
vascular	in	74 of	490
saline	in	72 of	490
acute	in	71 of	490
response	in	69 of	490
days	in	68 of	490
mm	in	68 of	490
dose	in	65 of	490
hg	in	62 of	490
activity	in	61 of	490
baseline	in	59 of	490
pulmonary	in	59 of	490
mmhg	in	58 of	490
values	in	58 of	490
tissue	in	57 of	490
heart	in	57 of	490
hypotension	in	56 of	490
fluid	in	55 of	490
septic	in	54 of	490
day	in	54 of	490
venous	in	53 of	490
oxygen	in	53 of	490
loss	in	53 of	490
solution	in	53 of	490
concentration	in	52 of	490
patient	in	52 of	490



ARMY Signal Processing Concept Exploration

Medline 16,000+



NEW Themescape - User Weighted View

weighted terms: resuscitation kg, plasma, oxygen

Appendix B

Definition of Variable Names

Definition of Variable Names

This appendix contains the complete list of variables calculated in the analysis for the **Signal Processing Concept Exploration** Task.

Table of Variable Names for All Patients

Note that variables 1, 2, and 3 were not used in the Exploratory Data Analysis. Variables 1 through 3 were the times at which the pattern detection algorithm identified the starting time for the heartbeat as shown in each of the three sensors, electrocardiogram (EKG), A-line, and Pulse Blood Oximeter, respectively.

Three analyses were performed for each patient: a simple, cross, and combination analysis. The variables used for each analysis are show below:

Mapping of Variables to Analysis Type

Type of Analysis	Variables Used
Simple Features	4 : 51
Cross Features	52: 195
Combination Features	4: 195

1 ekg time
2 aline time
3 blood o2 time
4 ekg diff
5 aline diff
6 blood o2 diff
7 R to Aline
8 R to Oxi
9 peak R
10 Aline amplitude
11 peak Aline
12 high d1 Aline
13 low d1 Aline
14 d1 width Aline
15 d2 width Aline
16 high d2 left Aline
17 high d2 right Aline
18 low d2 Aline
19 Oxi amplitude
20 trough Oxi
21 high d1 Oxi
22 low d1 Oxi
23 d1 width Oxi
24 d2 width Oxi
25 low d2 left Oxi
26 low d2 right Oxi
27 high d2 Oxi
28 integral Aline
29 integral Oximeter
30 aline ns Intercept
31 aline ns 1
32 aline ns 2

33 aline ns 3
34 aline ns 4
35 aline ns 5
36 aline ns 6
37 aline ns 7
38 aline ns 8
39 aline ns 9
40 aline ns 10
41 blood o2 ns
42 blood 02 ns 1
43 blood 02 ns 2
44 blood 02 ns 3
45 blood 02 ns 4
46 blood 02 ns 5
47 blood 02 ns 6
48 blood 02 ns 7
49 blood 02 ns 8
50 blood 02 ns 9
51 blood 02 ns 10
52 ekg diff mean 60 sec
53 ekg diff sd 60 sec
54 ekg diff skew 60 sec
55 aline diff mean 60 sec
56 aline diff sd 60 sec
57 aline diff skew 60 sec
58 blood o2 diff mean 60 sec
59 blood o2 diff sd 60 sec
60 blood o2 diff skew 60 sec
61 R to Aline mean 60 sec
62 R to Aline sd 60 sec
63 R to Aline skew 60 sec
64 R to Oxi mean 60 sec
65 R to Oxi sd 60 sec
66 R to Oxi skew 60 sec
67 peak R mean 60 sec
68 peak R sd 60 sec
69 peak R skew 60 sec
70 Aline amplitude mean 60 sec
71 Aline amplitude sd 60 sec
72 Aline amplitude skew 60 sec
73 peak Aline mean 60 sec
74 peak Aline sd 60 sec
75 peak Aline skew 60 sec
76 high d1 Aline mean 60 sec
77 high d1 Aline sd 60 sec
78 high d1 Aline skew 60 sec
79 low d1 Aline mean 60 sec
80 low d1 Aline sd 60 sec
81 low d1 Aline skew 60 sec
82 d1 width Aline mean 60 sec
83 d1 width Aline sd 60 sec
84 d1 width Aline skew 60 sec
85 d2 width Aline mean 60 sec
86 d2 width Aline sd 60 sec
87 d2 width Aline skew 60 sec
88 high d2 left Aline mean 60 sec

89 high d2 left Aline sd 60 sec
90 high d2 left Aline skew 60 sec
91 high d2 right Aline mean 60 sec
92 high d2 right Aline sd 60 sec
93 high d2 right Aline skew 60 sec
94 low d2 Aline mean 60 sec
95 low d2 Aline sd 60 sec
96 low d2 Aline skew 60 sec
97 Oxi amplitude mean 60 sec
98 Oxi amplitude sd 60 sec
99 Oxi amplitude skew 60 sec
100 trough Oxi mean 60 sec
101 trough Oxi sd 60 sec
102 trough Oxi skew 60 sec
103 high d1 Oxi mean 60 sec
104 high d1 Oxi sd 60 sec
105 high d1 Oxi skew 60 sec
106 low d1 Oxi mean 60 sec
107 low d1 Oxi sd 60 sec
108 low d1 Oxi skew 60 sec
109 d1 width Oxi mean 60 sec
110 d1 width Oxi sd 60 sec
111 d1 width Oxi skew 60 sec
112 d2 width Oxi mean 60 sec
113 d2 width Oxi sd 60 sec
114 d2 width Oxi skew 60 sec
115 low d2 left Oxi mean 60 sec
116 low d2 left Oxi sd 60 sec
117 low d2 left Oxi skew 60 sec
118 low d2 right Oxi mean 60 sec
119 low d2 right Oxi sd 60 sec
120 low d2 right Oxi skew 60 sec
121 high d2 Oxi mean 60 sec
122 high d2 Oxi sd 60 sec
123 high d2 Oxi skew 60 sec
124 integral Aline mean 60 sec
125 integral Aline sd 60 sec
126 integral Aline skew 60 sec
127 integral Oximeter mean 60 sec
128 integral Oximeter sd 60 sec
129 integral Oximeter skew 60 sec
130 aline ns Intercept mean 60 sec
131 aline ns Intercept sd 60 sec
132 aline ns Intercept skew 60 sec
133 aline ns 1 mean 60 sec
134 aline ns 1 sd 60 sec
135 aline ns 1 skew 60 sec
136 aline ns 2 mean 60 sec
137 aline ns 2 sd 60 sec
138 aline ns 2 skew 60 sec
139 aline ns 3 mean 60 sec
140 aline ns 3 sd 60 sec
141 aline ns 3 skew 60 sec
142 aline ns 4 mean 60 sec
143 aline ns 4 sd 60 sec
144 aline ns 4 skew 60 sec

145 aline ns 5 mean 60 sec
146 aline ns 5 sd 60 sec
147 aline ns 5 skew 60 sec
148 aline ns 6 mean 60 sec
149 aline ns 6 sd 60 sec
150 aline ns 6 skew 60 sec
151 aline ns 7 mean 60 sec
152 aline ns 7 sd 60 sec
153 aline ns 7 skew 60 sec
154 aline ns 8 mean 60 sec
155 aline ns 8 sd 60 sec
156 aline ns 8 skew 60 sec
157 aline ns 9 mean 60 sec
158 aline ns 9 sd 60 sec
159 aline ns 9 skew 60 sec
160 aline ns 10 mean 60 sec
161 aline ns 10 sd 60 sec
162 aline ns 10 skew 60 sec
163 blood o2 ns mean 60 sec
164 blood o2 ns sd 60 sec
165 blood o2 ns skew 60 sec
166 blood 02 ns 1 mean 60 sec
167 blood 02 ns 1 sd 60 sec
168 blood 02 ns 1 skew 60 sec
169 blood 02 ns 2 mean 60 sec
170 blood 02 ns 2 sd 60 sec
171 blood 02 ns 2 skew 60 sec
172 blood 02 ns 3 mean 60 sec
173 blood 02 ns 3 sd 60 sec
174 blood 02 ns 3 skew 60 sec
175 blood 02 ns 4 mean 60 sec
176 blood 02 ns 4 sd 60 sec
177 blood 02 ns 4 skew 60 sec
178 blood 02 ns 5 mean 60 sec
179 blood 02 ns 5 sd 60 sec
180 blood 02 ns 5 skew 60 sec
181 blood 02 ns 6 mean 60 sec
182 blood 02 ns 6 sd 60 sec
183 blood 02 ns 6 skew 60 sec
184 blood 02 ns 7 mean 60 sec
185 blood 02 ns 7 sd 60 sec
186 blood 02 ns 7 skew 60 sec
187 blood 02 ns 8 mean 60 sec
188 blood 02 ns 8 sd 60 sec
189 blood 02 ns 8 skew 60 sec
190 blood 02 ns 9 mean 60 sec
191 blood 02 ns 9 sd 60 sec
192 blood 02 ns 9 skew 60 sec
193 blood 02 ns 10 mean 60 sec
194 blood 02 ns 10 sd 60 sec
195 blood 02 ns 10 skew 60 sec

Appendix C

Simple Features – Pattern of Missing Values

Simple Features – Pattern of Missing Values

This appendix contains visualizations for time periods during which simple features are missing in the data record for a specific patient. If a system were built such as that described in Appendix G, "Medisense," it would be important to allow the user to view sensor data for which the pattern matcher failed to detect features and events. Any time both an EKG feature was identified *and* we were not able to include the variable shown at the left in the final analysis, a dot is drawn in the matrix at that time (horizontal axis) for that variable (vertical axis). Because of the high sampling frequency, most periods have some data missing.

The "Simple Feature ID" number shown on the far left in each of these plots is the variable number shown in the list found in Appendix B, "Definition of Variable Names." The lines at the top and bottom of the chart (at 0 and at 55) are summary lines and not for any specific variable. The line at y=0 shows where EKG features were identified. Thus gaps in this line indicate periods where no EKG-based heartbeat events were identified in the data. The line at y=55 aggregates all the missing information across variables for those times indicated by the line y=0.

The worst feature capture rate is for the integral variables, 28:29 and the donut coefficient variables 30:51.

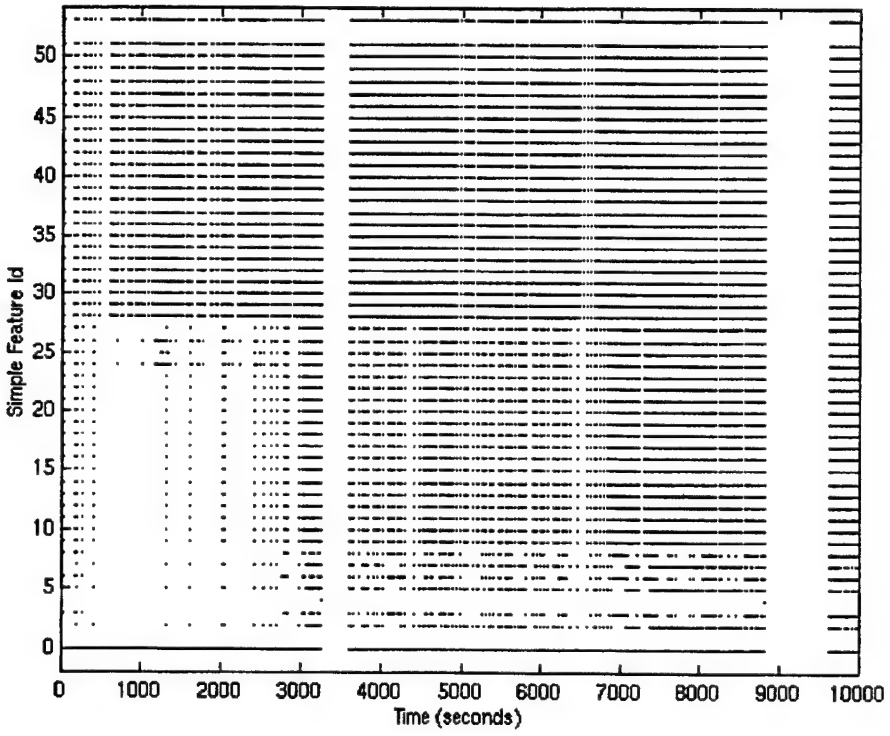


Figure C.1 Patient B -Pattern of Missing Values for the Simple Feature Calculations

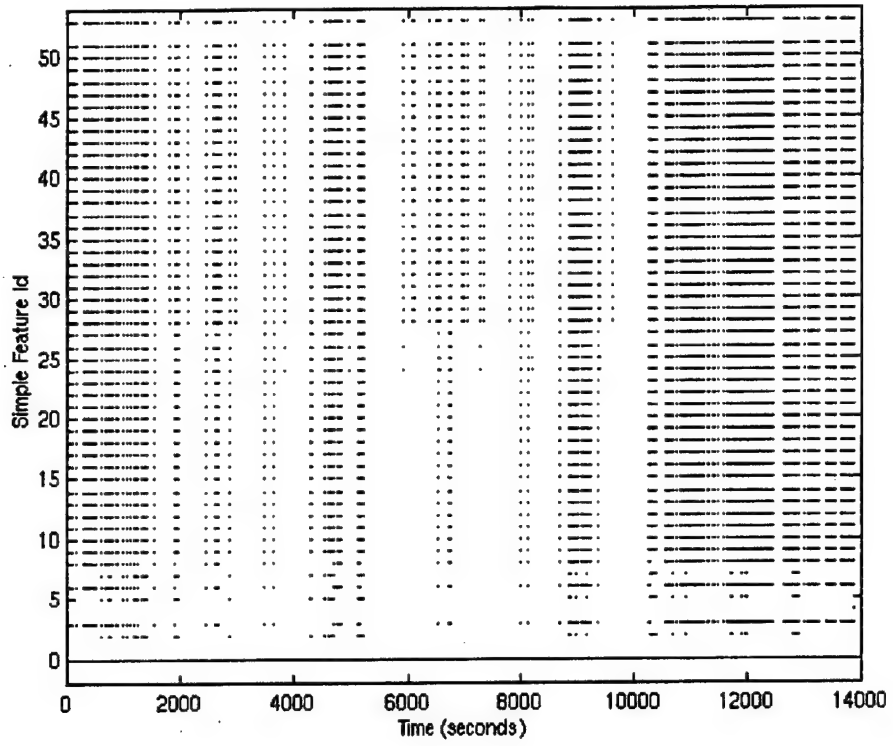


Figure C.2 – Patient H - Pattern of Missing Values the Simple Feature Calculations

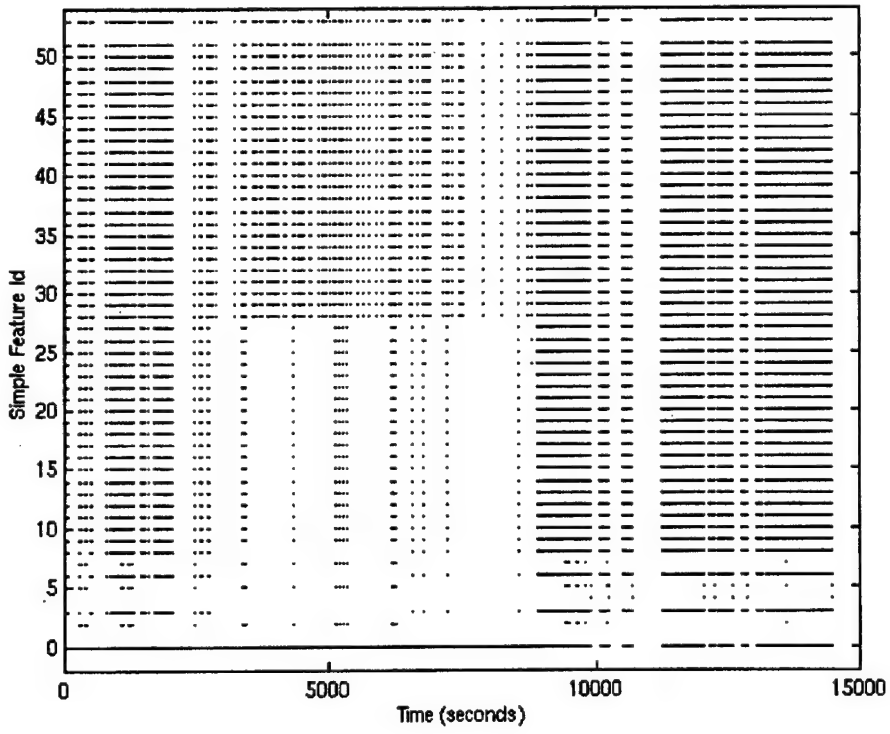


Figure C.3 – Patient W - Pattern of Missing Values for the Simple Feature Calculations

Appendix D

Cross Features – More Sample Calculations

Cross Features – More Sample Calculations

Appendix D shows visualizations for each patient for the calculation of the cross-variable features associated with the "Peak R" calculations. The last two panels are particularly important as the interplay between missing data and heart rate can be observed. Discounting for missing data, trends in heart rate can be observed. The pacemaker patient, H, is clearly differentiated from the other patients by the lower standard deviation in the cross Peak R feature calculations shown in panel 2 and in the pseudo-heart-rate plot shown in panel 4.Variables that were used in the cross and combined variable collections included only the mean, standard deviation, and skewness variables.

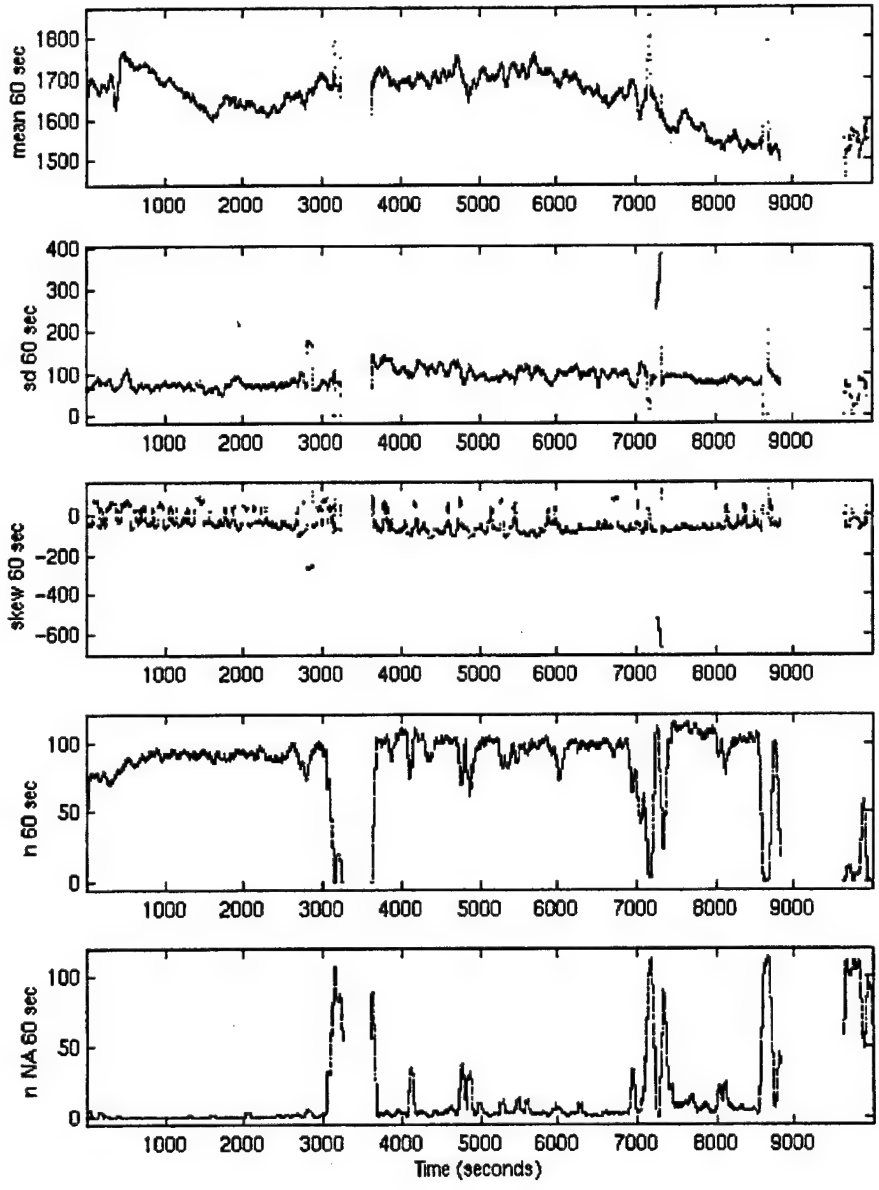


Figure D.1 - Patient B - Cross Peak R Calculations

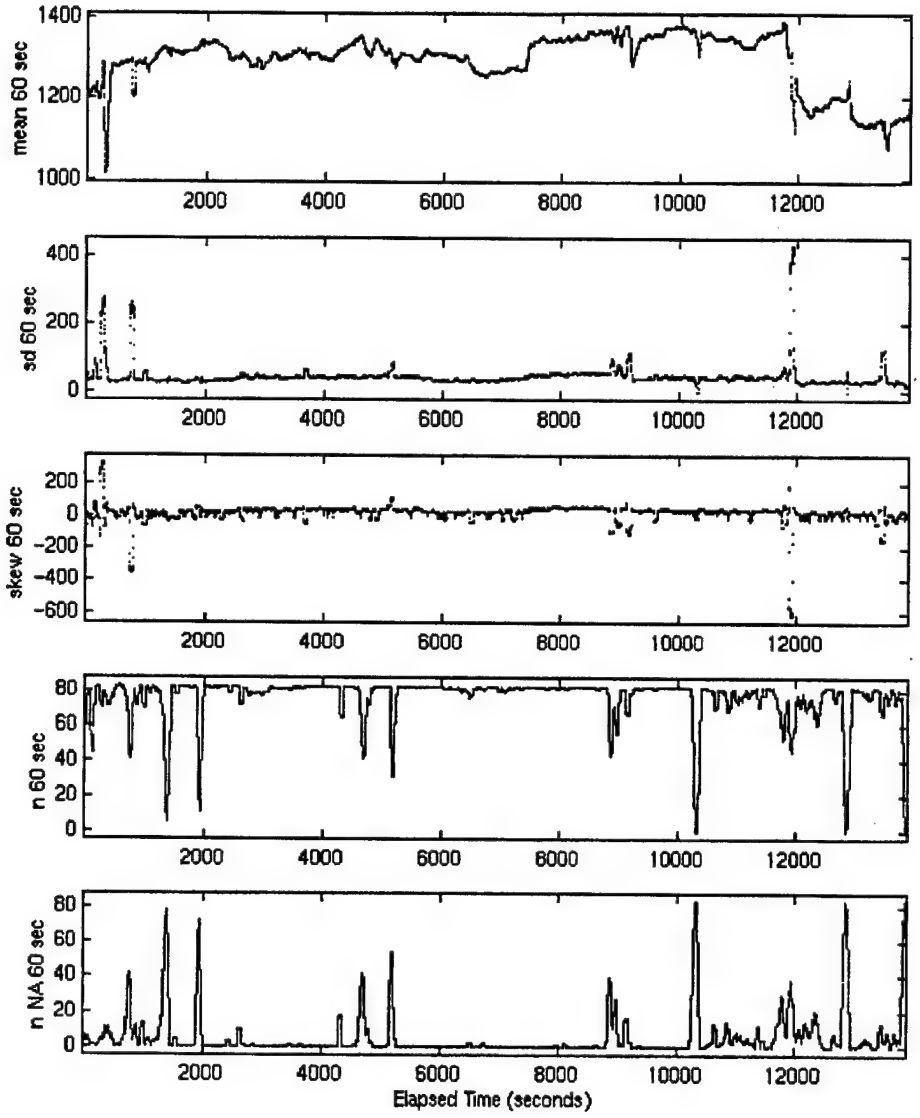


Figure D.2 - Patient H - Cross Peak R Calculations

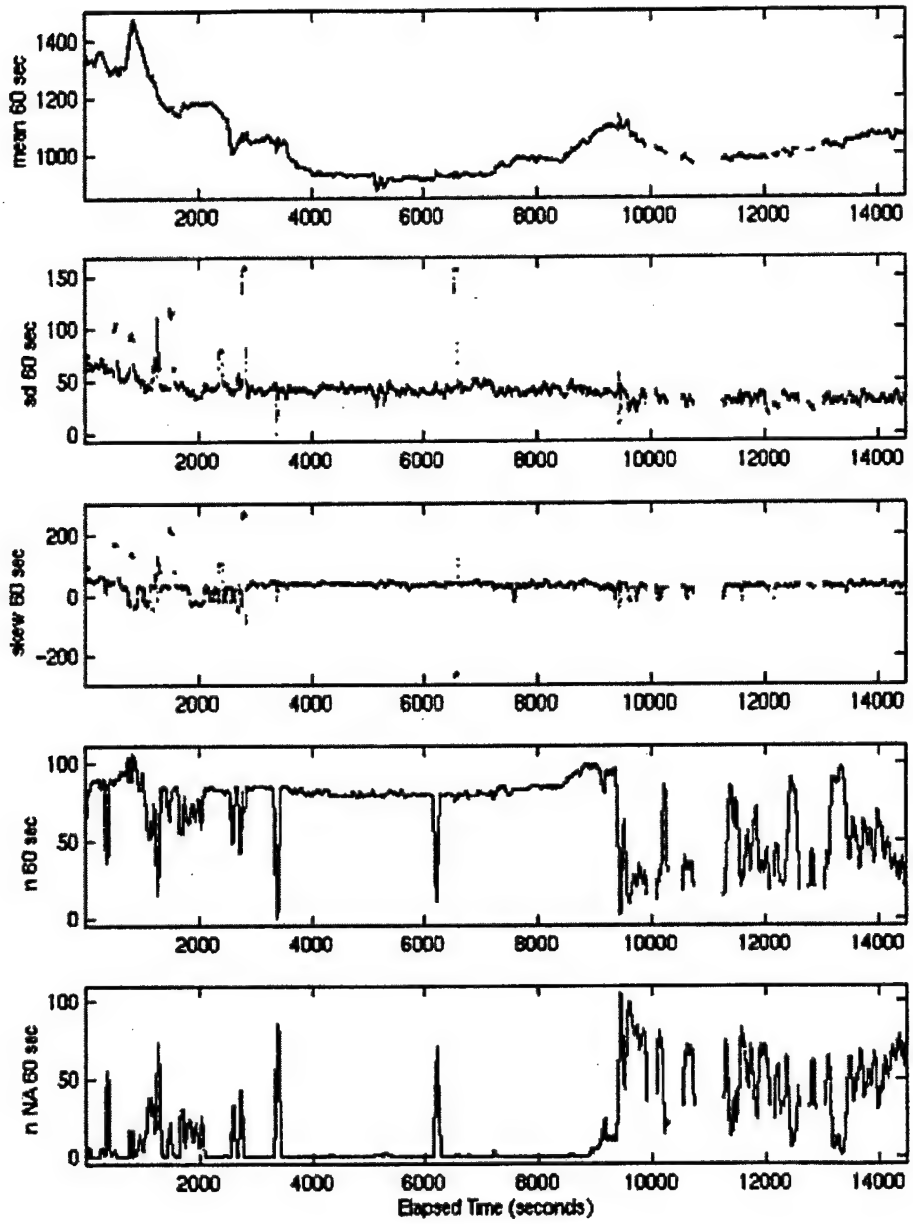


Figure D.3 - Patient W - Cross Peak R Calculations

Appendix E

Cross Features – Pattern of Missing Values

Cross Features – Pattern of Missing Values

This appendix contains visualizations for time periods during which cross features are missing in the data record for a specific patient. For comparison purposes, the visualizations of the simple features are also included at the bottom of each plot. If a system were built such as that described in Appendix G, "Medisense," it would be important to allow the user to view sensor data for which the pattern matcher failed to detect features and events. Any time both an electrocardiogram (EKG) feature was identified *and* we were not able to include the variable shown at the left in the final analysis, a dot is drawn in the matrix at that time (horizontal axis) for that variable (vertical axis). Because of the high sampling frequency, most periods have some data missing.

The "Feature ID" number shown on the far left in each of these plots is the variable number shown in the list found in Appendix B, "Definition of Variable Names." The lines at the top and bottom of the chart (at 0 and at 200) are summary lines and not for any specific variable. The line at y=0 shows where EKG features were identified. Thus, gaps in this line indicate periods where no EKG-based heartbeat events were identified in the data. The line at y=200 aggregates all the missing information across variables for those times indicated by the line y=0.

The feature capture rate for the integral and donut coefficients is greatly improved for the smoothed features.

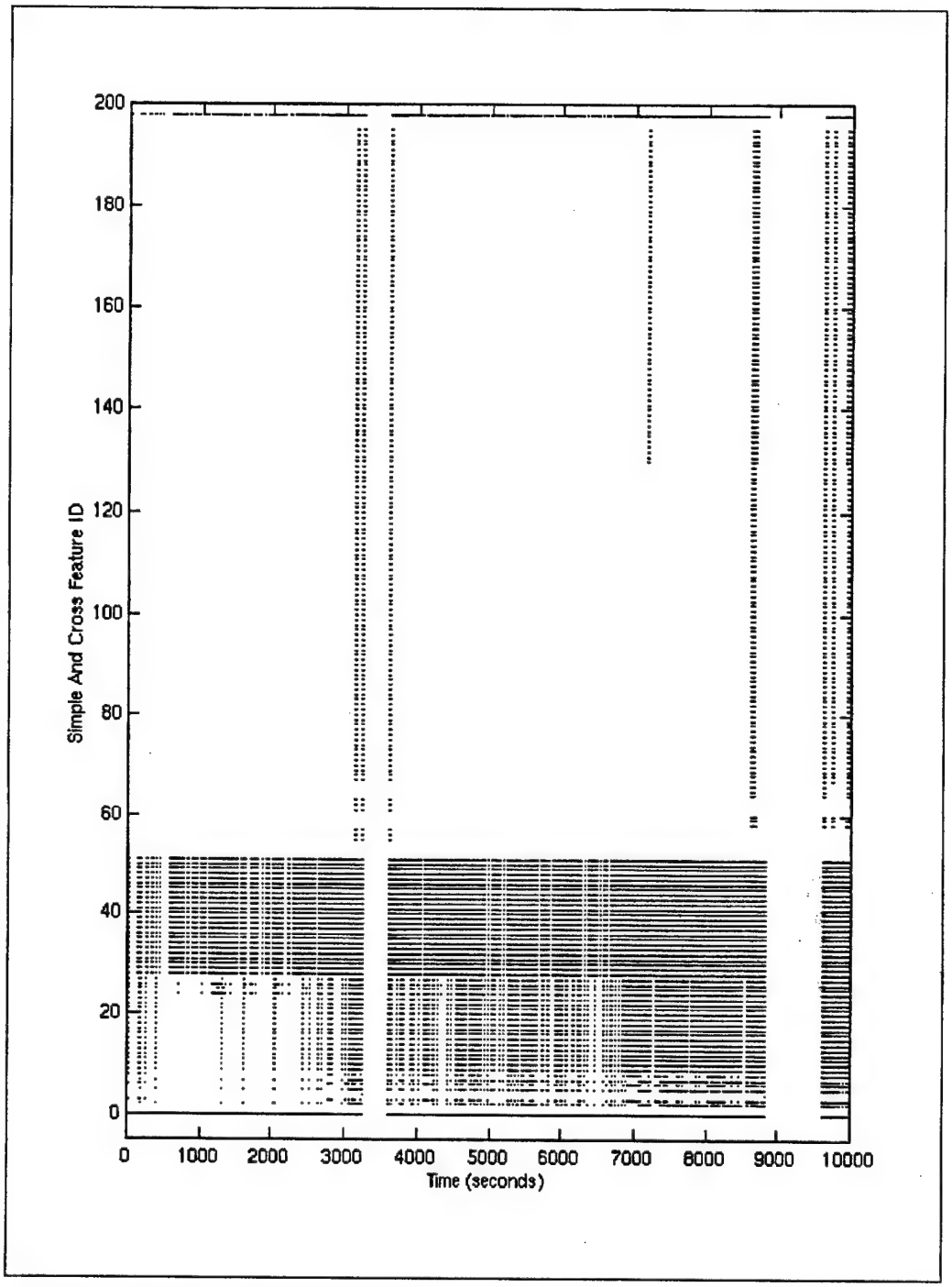


Figure E.1 – Patient B - Pattern of Missing Values for the Simple Feature Calculations

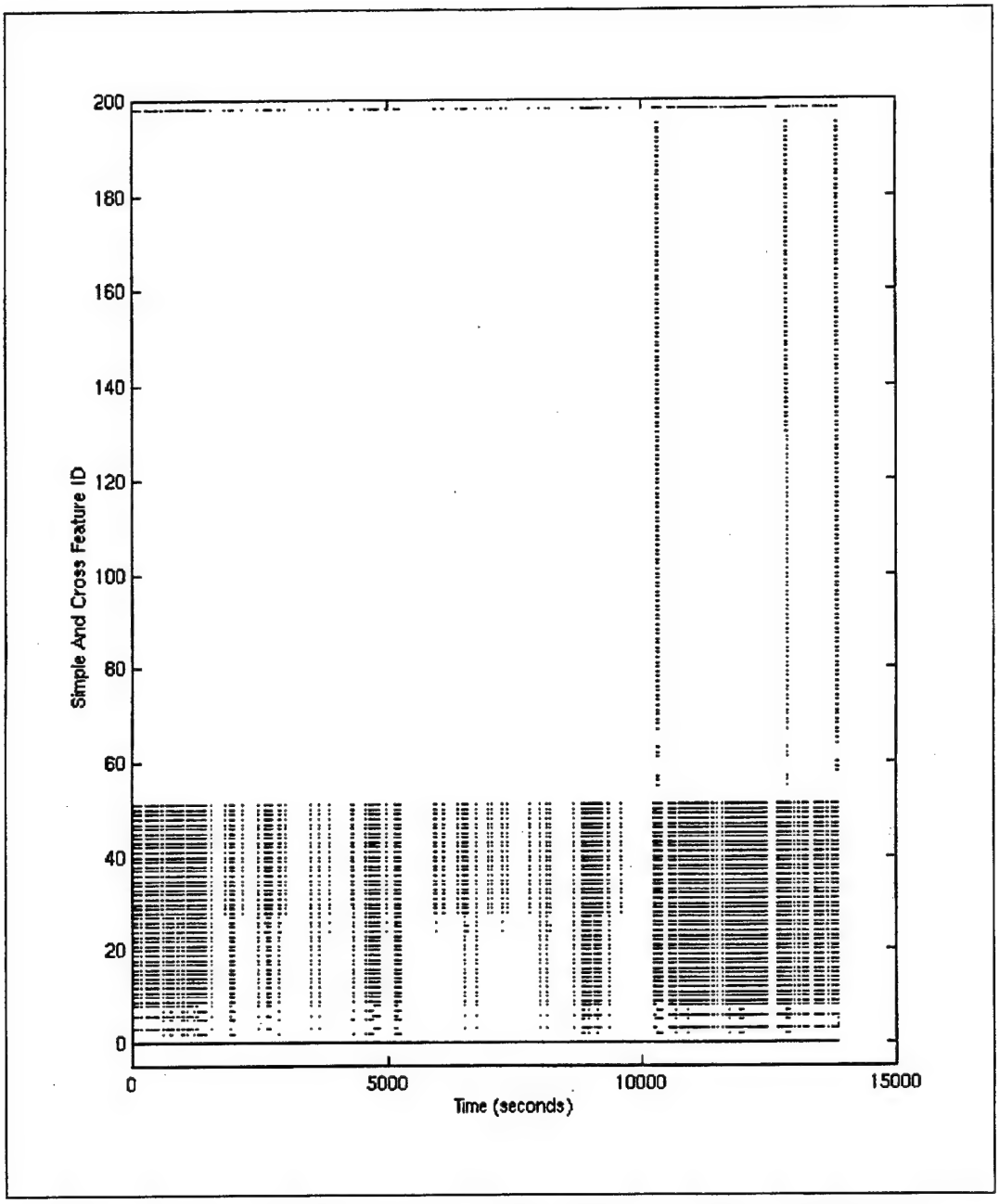


Figure E.2 - Patient H - Pattern of Missing Values for the Simple Feature Calculations

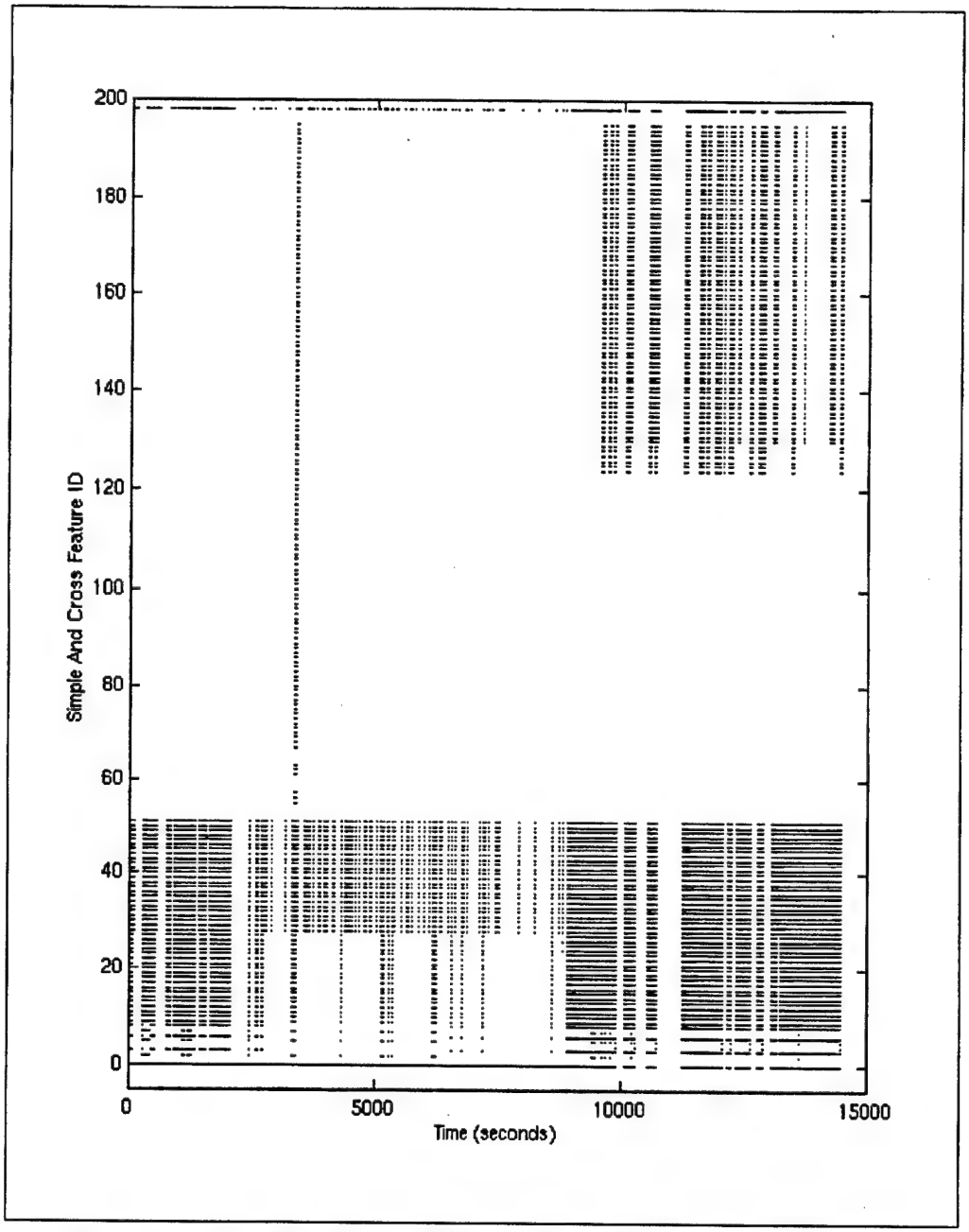


Figure E.3 – Patient W - Pattern of Missing Values for the Simple Feature Calculations

Appendix F

PCA Loadings

PCA Loadings

In this appendix, tables are provided for each patient by variable collection. These tables specify the 10 variables with the largest absolute value for their loading coefficients for the first six Principal Components (PCs). A summary plot for each of the six PCs is placed after each patient-variable table collection. The plot is an effective way to contrast the sign and relative magnitude for the loadings.

Patient B

Patient B Simple Feature Set

PC	Rank	ID	Feature
1	1	12	high d1 Aline
1	2	16	high d2 left Aline
1	3	10	Aline amplitude
1	4	18	low d2 Aline
1	5	11	peak Aline
1	6	28	integral Aline
1	7	7	R to Aline
1	8	13	low d1 Aline
1	9	48	blood 02 ns 7
1	10	17	high d2 right Aline

PC	Rank	ID	Feature
2	1	45	blood 02 ns 4
2	2	46	blood 02 ns 5
2	3	20	trough Oxi
2	4	41	blood o2 ns
2	5	44	blood 02 ns 3
2	6	19	Oxi amplitude
2	7	22	low d1 Oxi
2	8	5	aline diff
2	9	29	integral Oximeter
2	10	42	blood 02 ns 1

PC	Rank	ID	Feature
3	1	33	aline ns 3
3	2	49	blood 02 ns 8
3	3	48	blood 02 ns 7
3	4	36	aline ns 6
3	5	35	aline ns 5
3	6	51	blood 02 ns 10
3	7	50	blood 02 ns 9
3	8	6	blood o2 diff
3	9	4	ekg diff
3	10	11	peak Aline

PC	Rank	ID	Feature
4	1	42	blood 02 ns 1
4	2	21	high d1 Oxi
4	3	19	Oxi amplitude
4	4	43	blood 02 ns 2
4	5	50	blood 02 ns 9
4	6	44	blood 02 ns 3
4	7	27	high d2 Oxi
4	8	26	low d2 right Oxi
4	9	22	low d1 Oxi
4	10	29	integral Oximeter

PC	Rank	ID	Feature
5	1	37	aline ns 7
5	2	32	aline ns 2
5	3	39	aline ns 9
5	4	34	aline ns 4
5	5	31	aline ns 1
5	6	47	blood 02 ns 6
5	7	30	aline ns Intercept
5	8	38	aline ns 8
5	9	46	blood 02 ns 5
5	10	43	blood 02 ns 2

PC	Rank	ID	Feature
6	1	40	aline ns 10
6	2	38	aline ns 8
6	3	39	aline ns 9
6	4	30	aline ns Intercept
6	5	32	Aline ns 2
6	6	14	d1 width Aline
6	7	33	aline ns 3
6	8	34	aline ns 4
6	9	24	d2 width Oxi
6	10	15	d2 width Aline

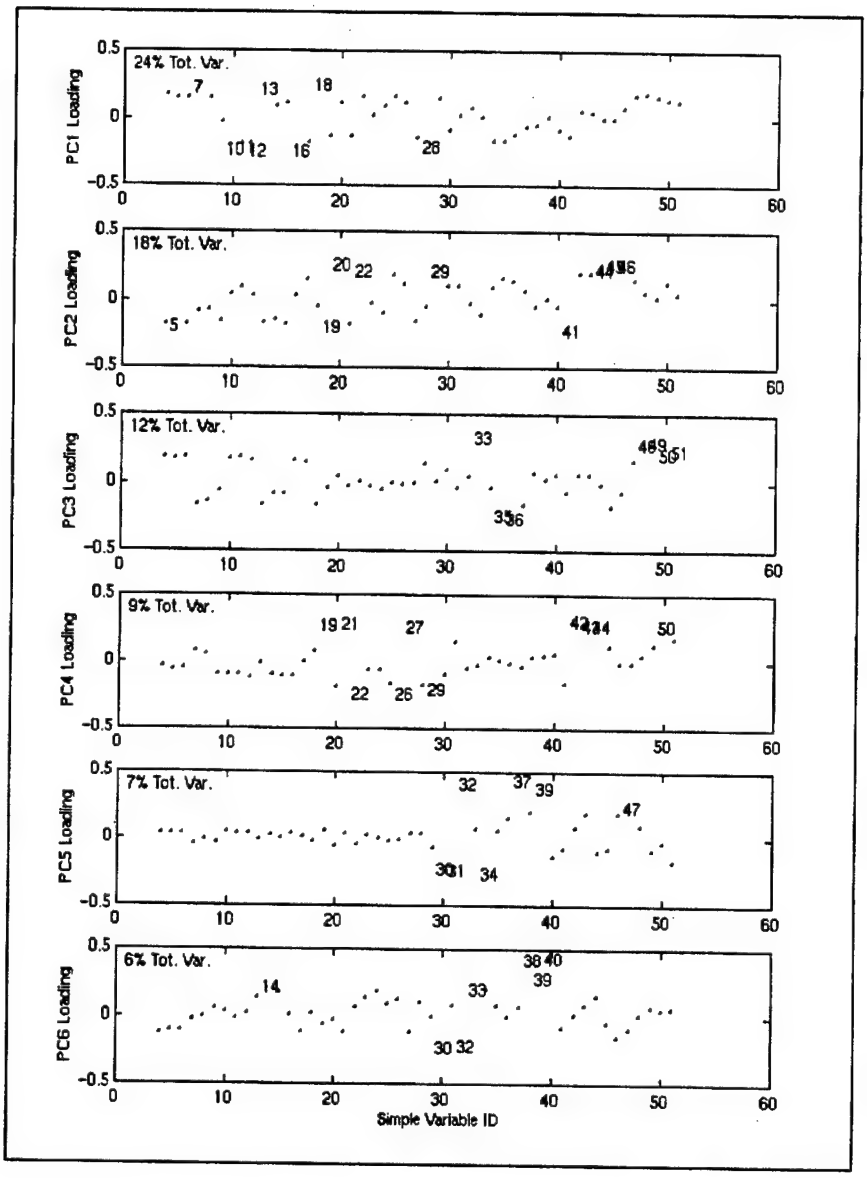


Figure F.1 - Patient B - Loadings for Simple Feature Set

Patient B Cross Feature Set

PC	Rank	ID	Feature
1	1	79	low d1 Aline mean 60 sec
1	2	77	high d1 Aline sd 60 sec
1	3	88	high d2 left Aline mean 60 sec
1	4	76	high d1 Aline mean 60 sec
1	5	85	d2 width Aline mean 60 sec
1	6	70	Aline amplitude mean 60 sec
1	7	149	aline ns 6 sd 60 sec
1	8	152	aline ns 7 sd 60 sec
1	9	145	aline ns 5 mean 60 sec
1	10	89	high d2 left Aline sd 60 sec

PC	Rank	ID	Feature
2	1	176	blood 02 ns 4 sd 60 sec
2	2	97	Oxi amplitude mean 60 sec
2	3	182	blood 02 ns 6 sd 60 sec
2	4	100	trough Oxi mean 60 sec
2	5	127	integral Oximeter mean 60 sec
2	6	185	blood 02 ns 7 sd 60 sec
2	7	179	blood 02 ns 5 sd 60 sec
2	8	106	low d1 Oxi mean 60 sec
2	9	173	blood 02 ns 3 sd 60 sec
2	10	101	trough Oxi sd 60 sec

PC	Rank	ID	Feature
3	1	163	blood o2 ns mean 60 sec
3	2	178	blood 02 ns 5 mean 60 sec
3	3	175	blood 02 ns 4 mean 60 sec
3	4	134	aline ns 1 sd 60 sec
3	5	172	blood 02 ns 3 mean 60 sec
3	6	181	blood 02 ns 6 mean 60 sec
3	7	169	blood 02 ns 2 mean 60 sec
3	8	190	blood 02 ns 9 mean 60 sec
3	9	158	aline ns 9 sd 60 sec
3	10	166	blood 02 ns 1 mean 60 sec

PC	Rank	ID	Feature
4	1	165	blood o2 ns skew 60 sec
4	2	129	integral Oximeter skew 60 sec
4	3	99	Oxi amplitude skew 60 sec
4	4	113	d2 width Oxi sd 60 sec
4	5	183	blood 02 ns 6 skew 60 sec
4	6	105	high d1 Oxi skew 60 sec
4	7	103	high d1 Oxi mean 60 sec
4	8	108	low d1 Oxi skew 60 sec
4	9	192	blood 02 ns 9 skew 60 sec
4	10	128	integral Oximeter sd 60 sec

PC	Rank	ID	Feature
5	1	75	peak Aline skew 60 sec
5	2	139	aline ns 3 mean 60 sec
5	3	141	aline ns 3 skew 60 sec
5	4	187	blood 02 ns 8 mean 60 sec
5	5	177	blood 02 ns 4 skew 60 sec
5	6	184	blood 02 ns 7 mean 60 sec
5	7	78	high d1 Aline skew 60 sec
5	8	193	blood 02 ns 10 mean 60 sec
5	9	195	blood 02 ns 10 skew 60 sec
5	10	133	aline ns 1 mean 60 sec

PC	Rank	ID	Feature
6	1	92	high d2 right Aline sd 60 sec
6	2	80	low d1 Aline sd 60 sec
6	3	83	d1 width Aline sd 60 sec
6	4	93	high d2 right Aline skew 60 sec
6	5	81	low d1 Aline skew 60 sec
6	6	84	d1 width Aline skew 60 sec
6	7	82	d1 width Aline mean 60 sec
6	8	193	blood 02 ns 10 mean 60 sec
6	9	190	blood 02 ns 9 mean 60 sec
6	10	187	blood 02 ns 8 mean 60 sec

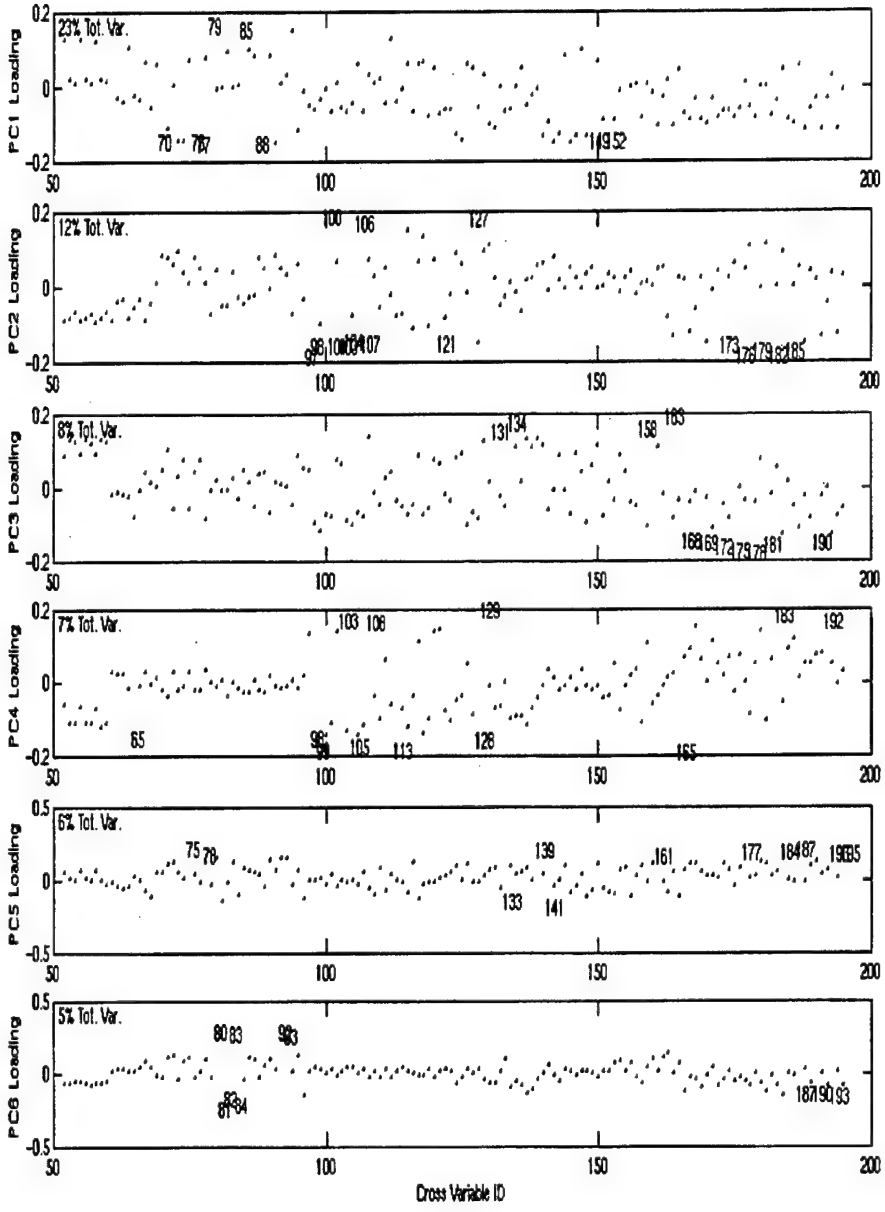


Figure F.2 - Patient B - Loadings For Cross Feature Set

Patient B Full Feature Set

PC	Rank	ID	Feature
1	1	79	low d1 Aline mean 60 sec
1	2	77	high d1 Aline sd 60 sec
1	3	149	aline ns 6 sd 60 sec
1	4	152	aline ns 7 sd 60 sec
1	5	13	low d1 Aline
1	6	88	high d2 left Aline mean 60 sec
1	7	142	aline ns 4 mean 60 sec
1	8	145	aline ns 5 mean 60 sec
1	9	94	low d2 Aline mean 60 sec
1	10	89	high d2 left Aline sd 60 sec

PC	Rank	ID	Feature
2	1	100	trough Oxi mean 60 sec
2	2	97	Oxi amplitude mean 60 sec
2	3	176	blood 02 ns 4 sd 60 sec
2	4	106	low d1 Oxi mean 60 sec
2	5	103	high d1 Oxi mean 60 sec
2	6	127	integral Oximeter mean 60 sec
2	7	182	blood 02 ns 6 sd 60 sec
2	8	121	high d2 Oxi mean 60 sec
2	9	185	blood 02 ns 7 sd 60 sec
2	10	115	low d2 left Oxi mean 60 sec

PC	Rank	ID	Feature
3	1	180	blood 02 ns 5 skew 60 sec
3	2	139	aline ns 3 mean 60 sec
3	3	175	blood 02 ns 4 mean 60 sec
3	4	108	low d1 Oxi skew 60 sec
3	5	150	aline ns 6 skew 60 sec
3	6	178	blood 02 ns 5 mean 60 sec
3	7	129	integral Oximeter skew 60 sec
3	8	134	aline ns 1 sd 60 sec
3	9	163	blood o2 ns mean 60 sec
3	10	99	Oxi amplitude skew 60 sec

PC	Rank	ID	Feature
4	1	113	d2 width Oxi sd 60 sec
4	2	99	Oxi amplitude skew 60 sec
4	3	128	integral Oximeter sd 60 sec
4	4	158	aline ns 9 sd 60 sec
4	5	134	aline ns 1 sd 60 sec
4	6	137	aline ns 2 sd 60 sec
4	7	129	integral Oximeter skew 60 sec
4	8	98	Oxi amplitude sd 60 sec
4	9	59	blood o2 diff sd 60 sec
4	10	108	low d1 Oxi skew 60 sec

PC	Rank	ID	Feature
5	1	50	blood 02 ns 9
5	2	42	blood 02 ns 1
5	3	41	blood o2 ns
5	4	48	blood 02 ns 7
5	5	47	blood 02 ns 6
5	6	43	blood 02 ns 2
5	7	28	integral Aline
5	8	49	blood 02 ns 8
5	9	51	blood 02 ns 10
5	10	46	blood 02 ns 5

PC	Rank	ID	Feature
6	1	49	blood 02 ns 8
6	2	60	blood o2 diff skew 60 sec
6	3	48	blood 02 ns 7
6	4	56	aline diff sd 60 sec
6	5	57	aline diff skew 60 sec
6	6	54	ekg diff skew 60 sec
6	7	59	blood o2 diff sd 60 sec
6	8	53	ekg diff sd 60 sec
6	9	36	aline ns 6
6	10	50	blood 02 ns 9

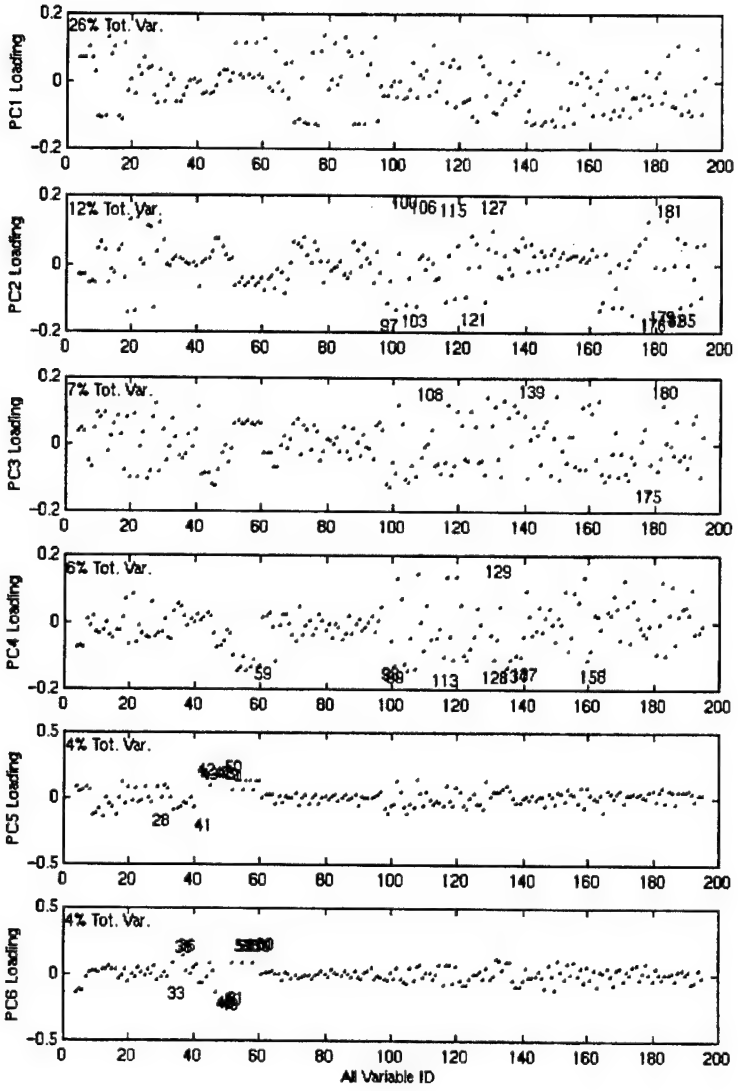


Figure F.3 - Patient B - Loadings for Full Feature Set

Patient H

Patient H Simple Feature Set

PC	Rank	ID	Feature
1	1	28	integral Aline
1	2	11	peak Aline
1	3	10	Aline amplitude
1	4	33	aline ns 3
1	5	12	high d1 Aline
1	6	34	aline ns 4
1	7	30	aline ns Intercept
1	8	32	aline ns 2
1	9	13	low d1 Aline
1	10	18	low d2 Aline

PC	Rank	ID	Feature
2	1	45	blood 02 ns 4
2	2	22	low d1 Oxi
2	3	20	trough Oxi
2	4	19	Oxi amplitude
2	5	46	blood 02 ns 5
2	6	21	high d1 Oxi
2	7	27	high d2 Oxi
2	8	41	blood o2 ns
2	9	44	blood 02 ns 3
2	10	29	integral Oximeter

PC	Rank	ID	Feature
3	1	49	blood 02 ns 8
3	2	48	blood 02 ns 7
3	3	50	blood 02 ns 9
3	4	51	blood 02 ns 10
3	5	47	blood 02 ns 6
3	6	41	blood 02 ns
3	7	21	high d1 Oxi
3	8	42	blood 02 ns 1
3	9	43	blood 02 ns 2
3	10	22	low d1 Oxi

PC	Rank	ID	Feature
4	1	37	aline ns 7
4	2	39	aline ns 9
4	3	5	aline diff
4	4	4	ekg diff
4	5	35	aline ns 5
4	6	36	aline ns 6
4	7	31	aline ns 1
4	8	38	aline ns 8
4	9	6	blood o2 diff
4	10	34	aline ns 4

PC	Rank	ID	Feature
5	1	6	blood o2 diff
5	2	8	R to Oxi
5	3	43	blood 02 ns 2
5	4	4	ekg diff
5	5	5	aline diff
5	6	42	blood 02 ns 1
5	7	16	high d2 left Aline
5	8	7	R to Aline
5	9	15	d2 width Aline
5	10	37	aline ns 7

PC	Rank	ID	Feature
6	1	8	R to Oxi
6	2	14	d1 width Aline
6	3	4	ekg diff
6	4	15	d2 width Aline
6	5	5	aline diff
6	6	24	d2 width Oxi
6	7	9	peak R
6	8	44	blood 02 ns 3
6	9	47	blood 02 ns 6
6	10	29	integral Oximeter

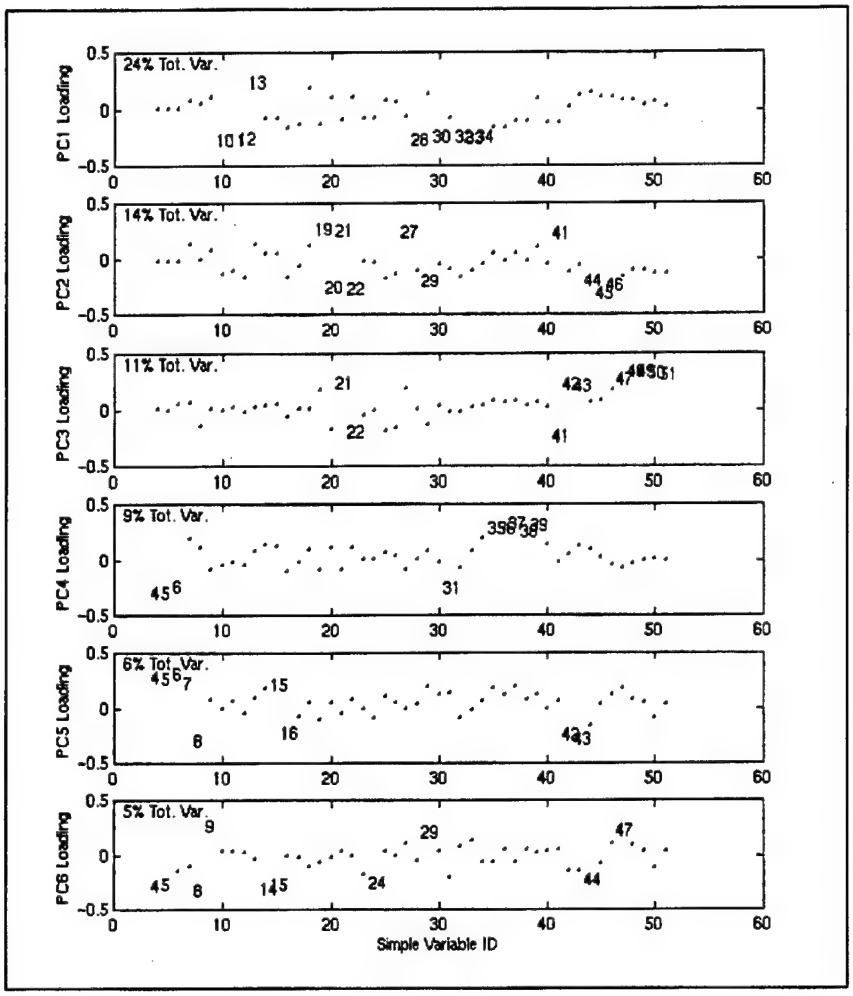


Figure F.4 - Patient H - Laodings For Simple Feature Set

Patient H Cross Feature Set

PC	Rank	ID	Feature
1	1	176	blood 02 ns 4 sd 60 sec
1	2	173	blood 02 ns 3 sd 60 sec
1	3	179	blood 02 ns 5 sd 60 sec
1	4	164	blood o2 ns sd 60 sec
1	5	188	blood 02 ns 8 sd 60 sec
1	6	191	blood 02 ns 9 sd 60 sec
1	7	194	blood 02 ns 10 sd 60 sec
1	8	185	blood 02 ns 7 sd 60 sec
1	9	170	blood 02 ns 2 sd 60 sec
1	10	182	blood 02 ns 6 sd 60 sec

PC	Rank	ID	Feature
2	1	172	blood 02 ns 3 mean 60 sec
2	2	163	blood o2 ns mean 60 sec
2	3	146	aline ns 5 sd 60 sec
2	4	143	aline ns 4 sd 60 sec
2	5	148	aline ns 6 mean 60 sec
2	6	154	aline ns 8 mean 60 sec
2	7	130	aline ns Intercept mean 60 sec
2	8	145	aline ns 5 mean 60 sec
2	9	137	aline ns 2 sd 60 sec
2	10	149	aline ns 6 sd 60 sec

PC	Rank	ID	Feature
3	1	136	aline ns 2 mean 60 sec
3	2	86	d2 width Aline sd 60 sec
3	3	88	high d2 left Aline mean 60 sec
3	4	76	high d1 Aline mean 60 sec
3	5	70	Aline amplitude mean 60 sec
3	6	79	low d1 Aline mean 60 sec
3	7	124	integral Aline mean 60 sec
3	8	157	aline ns 9 mean 60 sec
3	9	83	d1 width Aline sd 60 sec
3	10	139	aline ns 3 mean 60 sec

PC	Rank	ID	Feature
4	1	106	low d1 Oxi mean 60 sec
4	2	100	trough Oxi mean 60 sec
4	3	121	high d2 Oxi mean 60 sec
4	4	103	high d1 Oxi mean 60 sec
4	5	115	low d2 left Oxi mean 60 sec
4	6	175	blood 02 ns 4 mean 60 sec
4	7	97	Oxi amplitude mean 60 sec
4	8	127	integral Oximeter mean 60 sec
4	9	118	low d2 right Oxi mean 60 sec
4	10	178	blood 02 ns 5 mean 60 sec

PC	Rank	ID	Feature
5	1	56	aline diff sd 60 sec
5	2	57	aline diff skew 60 sec
5	3	54	ekg diff skew 60 sec
5	4	53	ekg diff sd 60 sec
5	5	59	blood o2 diff sd 60 sec
5	6	60	blood o2 diff skew 60 sec
5	7	55	aline diff mean 60 sec
5	8	52	ekg diff mean 60 sec
5	9	58	blood o2 diff mean 60 sec
5	10	177	blood 02 ns 4 skew 60 sec

PC	Rank	ID	Feature
6	1	132	aline ns Intercept skew 60 sec
6	2	153	aline ns 7 skew 60 sec
6	3	162	aline ns 10 skew 60 sec
6	4	159	aline ns 9 skew 60 sec
6	5	144	aline ns 4 skew 60 sec
6	6	156	aline ns 8 skew 60 sec
6	7	54	ekg diff skew 60 sec
6	8	150	aline ns 6 skew 60 sec
6	9	57	aline diff skew 60 sec
6	10	60	blood o2 diff skew 60 sec

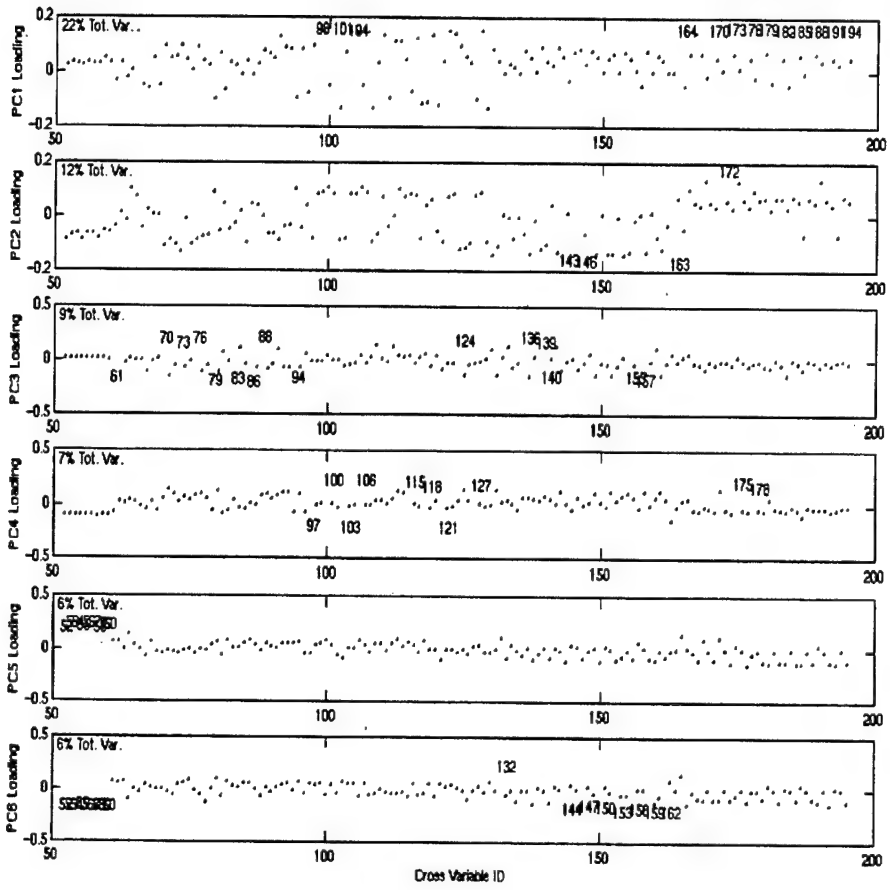


Figure F.5 - Patient H - Loadings for Cross Feature Set

Patient H Full Feature Set

PC	Rank	ID	Feature
1	1	91	high d2 right Aline mean 60 sec
1	2	124	integral Aline mean 60 sec
1	3	73	peak Aline mean 60 sec
1	4	11	peak Aline
1	5	28	integral Aline
1	6	79	low d1 Aline mean 60 sec
1	7	70	Aline amplitude mean 60 sec
1	8	142	aline ns 4 mean 60 sec
1	9	10	Aline amplitude
1	10	139	aline ns 3 mean 60 sec

PC	Rank	ID	Feature
2	1	172	blood 02 ns 3 mean 60 sec
2	2	163	blood o2 ns mean 60 sec
2	3	113	d2 width Oxi sd 60 sec
2	4	44	blood 02 ns 3
2	5	128	integral Oximeter sd 60 sec
2	6	108	low d1 Oxi skew 60 sec
2	7	107	low d1 Oxi sd 60 sec
2	8	98	Oxi amplitude sd 60 sec
2	9	129	integral Oximeter skew 60 sec
2	10	99	Oxi amplitude skew 60 sec

PC	Rank	ID	Feature
3	1	61	R to Aline mean 60 sec
3	2	88	high d2 left Aline mean 60 sec
3	3	136	aline ns 2 mean 60 sec
3	4	32	aline ns 2
3	5	157	aline ns 9 mean 60 sec
3	6	86	d2 width Aline sd 60 sec
3	7	16	high d2 left Aline
3	8	76	high d1 Aline mean 60 sec
3	9	12	high d1 Aline
3	10	160	aline ns 10 mean 60 sec

PC	Rank	ID	Feature
4	1	131	aline ns Intercept sd 60 sec
4	2	80	low d1 Aline sd 60 sec
4	3	161	aline ns 10 sd 60 sec
4	4	158	aline ns 9 sd 60 sec
4	5	149	aline ns 6 sd 60 sec
4	6	155	aline ns 8 sd 60 sec
4	7	152	aline ns 7 sd 60 sec
4	8	121	high d2 Oxi mean 60 sec
4	9	106	low d1 Oxi mean 60 sec
4	10	103	high d1 Oxi mean 60 sec

PC	Rank	ID	Feature
5	1	56	aline diff sd 60 sec
5	2	57	aline diff skew 60 sec
5	3	53	ekg diff sd 60 sec
5	4	54	ekg diff skew 60 sec
5	5	60	blood o2 diff skew 60 sec
5	6	59	blood o2 diff sd 60 sec
5	7	181	blood 02 ns 6 mean 60 sec
5	8	55	aline diff mean 60 sec
5	9	52	ekg diff mean 60 sec
5	10	58	blood o2 diff mean 60 sec

PC	Rank	ID	Feature
6	1	151	aline ns 7 mean 60 sec
6	2	177	blood 02 ns 4 skew 60 sec
6	3	134	aline ns 1 sd 60 sec
6	4	133	aline ns 1 mean 60 sec
6	5	165	blood o2 ns skew 60 sec
6	6	61	R to Aline mean 60 sec
6	7	174	blood 02 ns 3 skew 60 sec
6	8	192	blood 02 ns 9 skew 60 sec
6	9	145	aline ns 5 mean 60 sec
6	10	140	aline ns 3 sd 60 sec

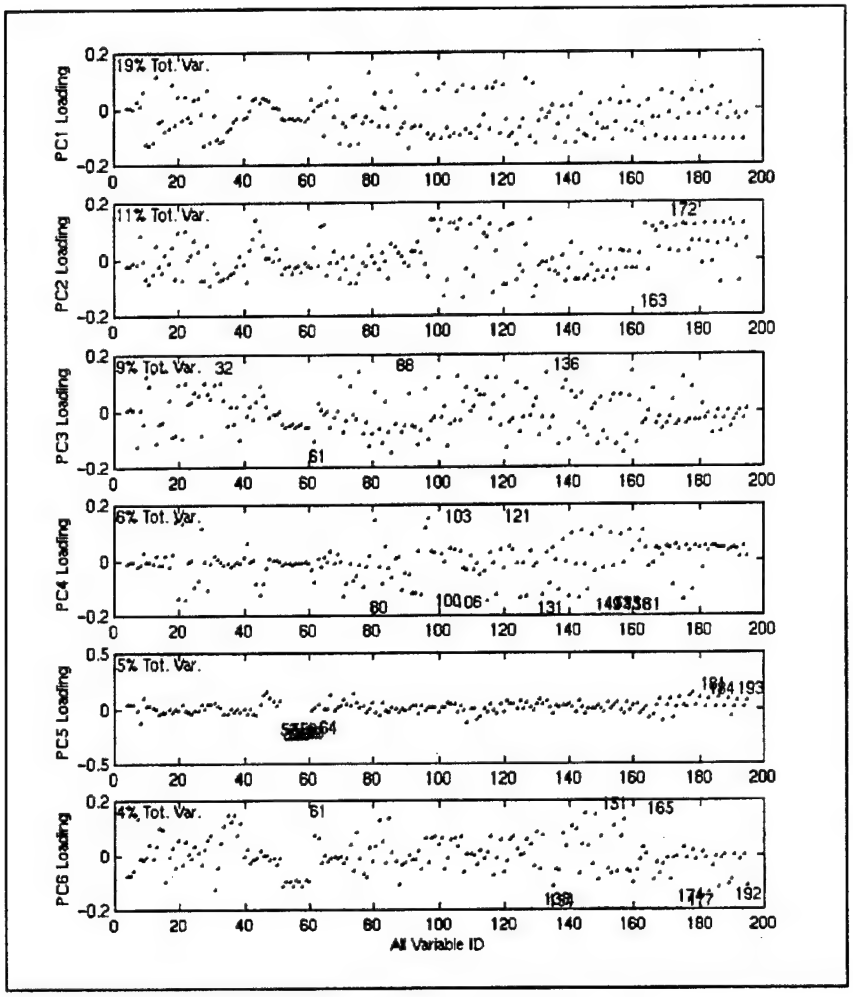


Figure F.6 - Patient H - Loadings for Full Feature Set

Patient W

Patient W Simple Feature Set

PC	Rank	ID	Feature
1	1	12	high d1 Aline
1	2	18	low d2 Aline
1	3	10	Aline amplitude
1	4	4	ekg diff
1	5	5	aline diff
1	6	16	high d2 left Aline
1	7	7	R to Aline
1	8	33	aline ns 3
1	9	6	blood o2 diff
1	10	11	peak Aline

PC	Rank	ID	Feature
2	1	21	high d1 Oxi
2	2	26	low d2 right Oxi
2	3	47	blood 02 ns 6
2	4	27	high d2 Oxi
2	5	22	low d1 Oxi
2	6	25	low d2 left Oxi
2	7	19	Oxi amplitude
2	8	35	aline ns 5
2	9	36	aline ns 6
2	10	45	blood 02 ns 4

PC	Rank	ID	Feature
3	1	46	blood 02 ns 5
3	2	44	blood 02 ns 3
3	3	41	blood o2 ns
3	4	45	blood 02 ns 4
3	5	29	integral Oximeter
3	6	20	trough Oxi
3	7	28	integral Aline
3	8	22	low d1 Oxi
3	9	19	Oxi amplitude
3	10	43	blood 02 ns 2

PC	Rank	ID	Feature
4	1	49	blood 02 ns 8
4	2	51	blood 02 ns 10
4	3	50	blood 02 ns 9
4	4	34	aline ns 4
4	5	42	blood 02 ns 1
4	6	47	blood 02 ns 6
4	7	15	d2 width Aline
4	8	41	blood o2 ns
4	9	28	integral Aline
4	10	30	aline ns Intercept

PC	Rank	ID	Feature
5	1	39	aline ns 9
5	2	38	aline ns 8
5	3	40	aline ns 10
5	4	42	blood 02 ns 1
5	5	36	aline ns 6
5	6	30	aline ns Intercept
5	7	37	aline ns 7
5	8	35	aline ns 5
5	9	34	aline ns 4
5	10	17	high d2 right Aline

PC	Rank	ID	Feature
6	1	46	blood 02 ns 5
6	2	14	d1 width Aline
6	3	17	high d2 right Aline
6	4	40	aline ns 10
6	5	24	d2 width Oxi
6	6	36	aline ns 6
6	7	34	aline ns 4
6	8	32	aline ns 2
6	9	15	d2 width Aline
6	10	51	blood 02 ns 10

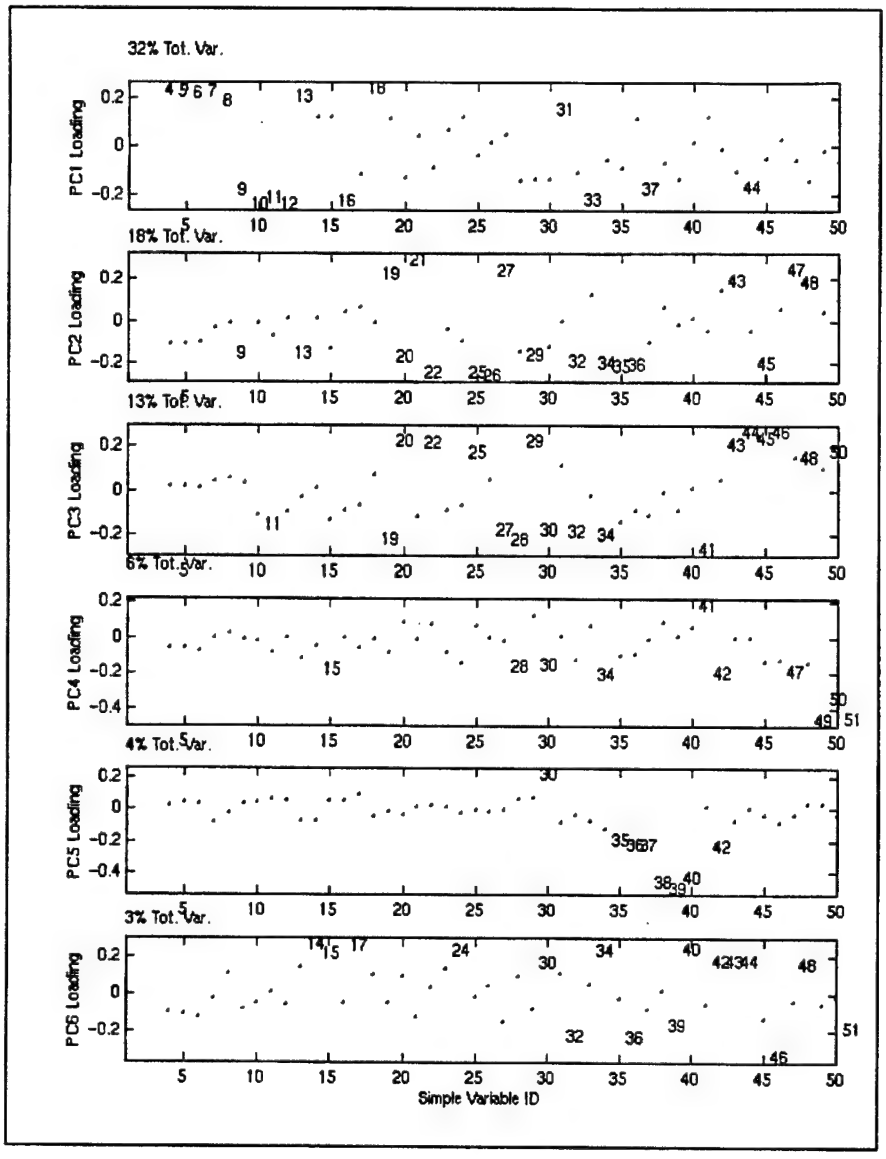


Figure F.7 - Patient W - Simple Feature Set

Patient W Cross Feature Set

PC	Rank	ID	Feature
1	1	118	low d2 right Oxi mean 60 sec
1	2	103	high d1 Oxi mean 60 sec
1	3	181	blood 02 ns 6 mean 60 sec
1	4	113	d2 width Oxi sd 60 sec
1	5	128	integral Oximeter sd 60 sec
1	6	86	d2 width Aline sd 60 sec
1	7	123	high d2 Oxi skew 60 sec
1	8	79	low d1 Aline mean 60 sec
1	9	125	integral Aline sd 60 sec
1	10	129	integral Oximeter skew 60 sec

PC	Rank	ID	Feature
2	1	61	R to Aline mean 60 sec
2	2	94	low d2 Aline mean 60 sec
2	3	76	high d1 Aline mean 60 sec
2	4	70	Aline amplitude mean 60 sec
2	5	52	ekg diff mean 60 sec
2	6	55	aline diff mean 60 sec
2	7	58	blood o2 diff mean 60 sec
2	8	88	high d2 left Aline mean 60 sec
2	9	73	peak Aline mean 60 sec
2	10	79	low d1 Aline mean 60 sec

PC	Rank	ID	Feature
3	1	146	aline ns 5 sd 60 sec
3	2	140	aline ns 3 sd 60 sec
3	3	155	aline ns 8 sd 60 sec
3	4	143	aline ns 4 sd 60 sec
3	5	149	aline ns 6 sd 60 sec
3	6	152	aline ns 7 sd 60 sec
3	7	137	aline ns 2 sd 60 sec
3	8	71	Aline amplitude sd 60 sec
3	9	158	aline ns 9 sd 60 sec
3	10	80	low d1 Aline sd 60 sec

PC	Rank	ID	Feature
4	1	175	blood 02 ns 4 mean 60 sec
4	2	166	blood 02 ns 1 mean 60 sec
4	3	127	integral Oximeter mean 60 sec
4	4	172	blood 02 ns 3 mean 60 sec
4	5	106	low d1 Oxi mean 60 sec
4	6	121	high d2 Oxi mean 60 sec
4	7	115	low d2 left Oxi mean 60 sec
4	8	195	blood 02 ns 10 skew 60 sec
4	9	133	aline ns 1 mean 60 sec
4	10	157	aline ns 9 mean 60 sec

PC	Rank	ID	Feature
5	1	174	blood 02 ns 3 skew 60 sec
5	2	177	blood 02 ns 4 skew 60 sec
5	3	180	blood 02 ns 5 skew 60 sec
5	4	165	blood o2 ns skew 60 sec
5	5	59	blood o2 diff sd 60 sec
5	6	186	blood 02 ns 7 skew 60 sec
5	7	136	aline ns 2 mean 60 sec
5	8	195	blood 02 ns 10 skew 60 sec
5	9	171	blood 02 ns 2 skew 60 sec
5	10	189	blood 02 ns 8 skew 60 sec

PC	Rank	ID	Feature
6	1	53	ekg diff sd 60 sec
6	2	56	aline diff sd 60 sec
6	3	54	ekg diff skew 60 sec
6	4	57	aline diff skew 60 sec
6	5	60	blood o2 diff skew 60 sec
6	6	144	aline ns 4 skew 60 sec
6	7	59	blood o2 diff sd 60 sec
6	8	138	aline ns 2 skew 60 sec
6	9	126	integral Aline skew 60 sec
6	10	153	aline ns 7 skew 60 sec

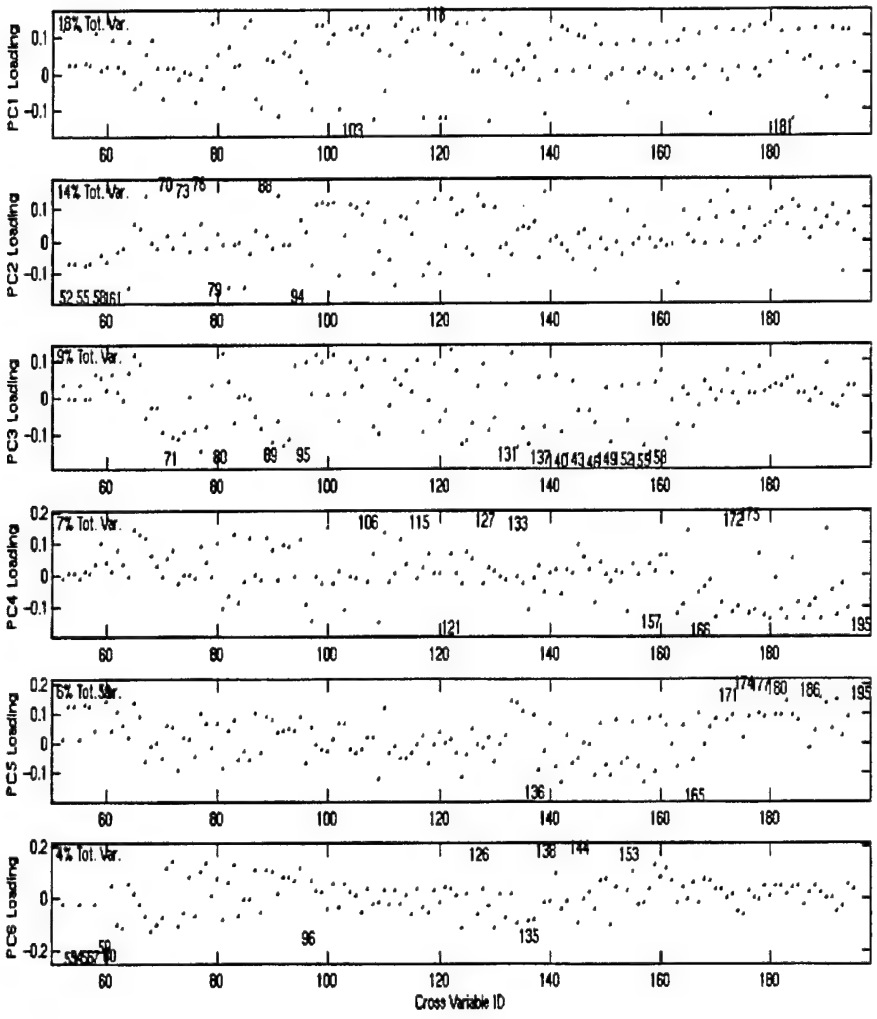


Figure F.8 - Patient W - Loadings for Cross Feature Set

Patient W Full Feature Set

PC	Rank	ID	Feature
1	1	94	low d2 Aline mean 60 sec
1	2	76	high d1 Aline mean 60 sec
1	3	70	Aline amplitude mean 60 sec
1	4	61	R to Aline mean 60 sec
1	5	18	low d2 Aline
1	6	12	high d1 Aline
1	7	88	high d2 left Aline mean 60 sec
1	8	10	Aline amplitude
1	9	64	R to Oxi mean 60 sec
1	10	52	ekg diff mean 60 sec

PC	Rank	ID	Feature
2	1	118	low d2 right Oxi mean 60 sec
2	2	181	blood 02 ns 6 mean 60 sec
2	3	103	high d1 Oxi mean 60 sec
2	4	113	d2 width Oxi sd 60 sec
2	5	21	high d1 Oxi
2	6	86	d2 width Aline sd 60 sec
2	7	148	aline ns 6 mean 60 sec
2	8	175	blood 02 ns 4 mean 60 sec
2	9	115	low d2 left Oxi mean 60 sec
2	10	121	high d2 Oxi mean 60 sec

PC	Rank	ID	Feature
3	1	143	aline ns 4 sd 60 sec
3	2	131	aline ns Intercept sd 60 sec
3	3	124	integral Aline mean 60 sec
3	4	125	integral Aline sd 60 sec
3	5	80	low d1 Aline sd 60 sec
3	6	101	trough Oxi sd 60 sec
3	7	146	aline ns 5 sd 60 sec
3	8	28	integral Aline
3	9	89	high d2 left Aline sd 60 sec
3	10	178	blood 02 ns 5 mean 60 sec

PC	Rank	ID	Feature
4	1	127	integral Oximeter mean 60 sec
4	2	97	Oxi amplitude mean 60 sec
4	3	115	low d2 left Oxi mean 60 sec
4	4	106	low d1 Oxi mean 60 sec
4	5	178	blood 02 ns 5 mean 60 sec
4	6	29	integral Oximeter
4	7	163	blood o2 ns mean 60 sec
4	8	19	Oxi amplitude
4	9	100	trough Oxi mean 60 sec
4	10	22	low d1 Oxi

PC	Rank	ID	Feature
5	1	155	aline ns 8 sd 60 sec
5	2	77	high d1 Aline sd 60 sec
5	3	152	aline ns 7 sd 60 sec
5	4	161	aline ns 10 sd 60 sec
5	5	95	low d2 Aline sd 60 sec
5	6	89	high d2 left Aline sd 60 sec
5	7	158	aline ns 9 sd 60 sec
5	8	81	low d1 Aline skew 60 sec
5	9	149	aline ns 6 sd 60 sec
5	10	71	Aline amplitude sd 60 sec

PC	Rank	ID	Feature
6	1	59	blood o2 diff sd 60 sec
6	2	53	ekg diff sd 60 sec
6	3	56	aline diff sd 60 sec
6	4	54	ekg diff skew 60 sec
6	5	60	blood o2 diff skew 60 sec
6	6	57	aline diff skew 60 sec
6	7	174	blood 02 ns 3 skew 60 sec
6	8	63	R to Aline skew 60 sec
6	9	62	R to Aline sd 60 sec
6	10	135	aline ns 1 skew 60 sec

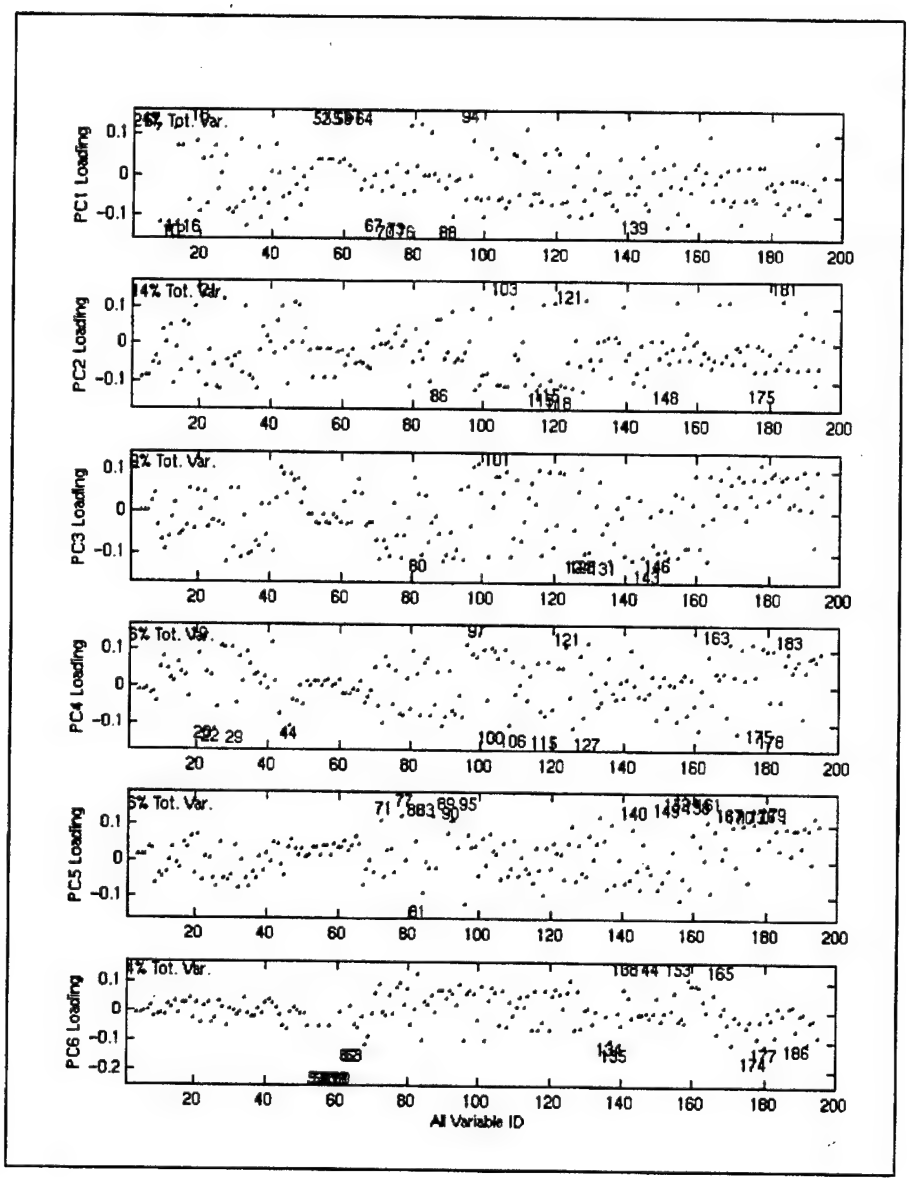


Figure F.9 - Patient W - Loadings for Full Feature Set

Appendix G

MediSense Proposal

MediSense Proposal
Signal Processing Concept Exploration and Display

G.1 Background

Multi-sensor medical data from patients are increasingly available. An example is the Life Support for Trauma and Transport (LSTAT) system that will house a variety of cardiac and respiratory sensors. The data from these sensors are potentially useful for real-time monitoring of patient conditions, diagnosing chronic health problems, and triage. However, because of the size and complexity of the data, it is difficult to survey even a few hours of sensor data, much less use the data for the three items above. Based on last fiscal year's work, we think that this project will take a giant step forward in examining these types of data.

This work is a continuation and enhancement of the Signal Processing Concept Exploration (SPCE) Task performed by the Pacific Northwest National Laboratory (PNNL) that is part of the Information Analysis and Visualization Research and Development Project for the U.S. Army Medical Research and Material Command (MRMC). Our clients are Major Stephen Bruttig of MRMC and Dr. Frederick Pearce of the Walter Reed Army Institute of Research Division of Surgery (WRAIR Surgery Division). Based on work in fiscal year 1997 (FY97) and the large amounts of data that the LSTAT project will generate, it may be more appropriate for this task to become its own project. The system we propose to develop is called ***MediSense***. Aspects of the envisioned system are "SPIRE-like" in that the data analysis is unsupervised and high-level features can be automatically extracted by the system, but the SPIRE system is not used.

G.2 Goals

The long-term goals of this work are to combine the data from a variety of real-time medical sensors from a single patient or from multiple patients so that

- a patient's detailed condition can be readily ascertained by medical care providers
- a patient's sensor history can be quickly reviewed by the medical care providers
- multiple patients' physical states can be assessed and compared, to prioritize and group the patients for efficient treatment of injuries.

The targeted application is the multiple sensors of the LSTAT stretcher.

G.3 Technical Approach

To achieve the above goals, this year's work will focus on the following:

1) **Detailed analysis and presentation of a patient's condition.**
This task examines LSTAT-like data from several patients at the heartbeat-to-heartbeat level of detail, over a specified time period of several hours or more. This work is a continuation of work begun in FY97. Activities that are included in this work are

   a) Calculating summaries and "low-order" features from the individual sensors. For instance, this year we developed algorithms to detect individual heartbeats in three sensors and then calculated summaries related to the "shape," spacing, and location of the heartbeat features.
   b) Detecting events in individual sensor data streams. For instance, this year we detected the absence of a clean heartbeat feature when such a feature was expected.
   c) Reducing dimension/data. Ways to view a few hours of a patient's sensor data, based on the summaries and events described in parts a) and b), will be explored. Alternatives will be evaluated and presented. A possibility for the data reduction is organizing for each patient the collection of features from the multiple sensors into an event feature matrix where time is one dimension and

derived features comprise another dimension. A statistical analysis, such as principal component analysis, is used to reduce the dimensionality of the event space to two or three. This allows the events, which are chunks of time from the patient sensor records, to be projected into "diagnostic-space."

2) **Visualization of the patient's current state.**
We will create prototypes of displays that can be used to summarize (in real time) a patient's condition. These displays may be a simple icon, glyph or other graphic. The displays will be based on the features derived in Activity 1. Simple interactions (designed to reveal more information as needed by the attending medical personnel) with the displays will also be prototyped.

3) **Cross-patient compressed views.**
This activity provides for a cross-patient visualization that spatially organizes the individual icons in "diagnostic or alarm" space. Several views are possible. A geometric-based view places the summary icon for a patient in a location on the screen suggested by where the patient is physically located. A diagnostic-based view places patients whose conditions are most similar close together on the screen. In any case, various conditions noted in the sensor data would signal the icon to go into alarm mode.

4) **Architectural issues for the MediSense system.**
This activity plans for the future by providing broad software and hardware architectural guidance for the **MediSense** analysis platform as it will apply to LSTAT (or related) sensor data. Getting this activity underway during this fiscal year will allow us to better leverage our exploratory work in the construction of the eventual product.

5) **Handling dropped sensors or other missing/corrupt data.**
It is inevitable that either a sensor(s) will fail or the sensor suite in LSTAT will evolve. This activity will explore how such inevitabilities will be handled; in particular, for many patient states, it may be that useful monitoring or triage decisions can be made with some subset of the full suite of sensors.

6) **Relationships between patient state (e.g., satisfactory, dying etc.) and sensor features.**
The relations between patient states and features derived from the sensors must be explored. For instance, what sensor states correspond with an individual who can be safely left alone for an hour or two? What sensor states correspond with an individual that is in immediate need of active care? What sensor states correspond with an individual who is beyond medical science's ability to resuscitate? These relationships between patient states and sensor measurements must be made for MediSense to be a diagnostic system (however, exploratory analysis of the data in Activity 1 does not require such relations to be known). This activity will gather together readily available relations and provide a plan for how this necessary relationship can be further elucidated. Note that standard options for obtaining such a relationship range from using available observations (e.g., we can use existing data to learn what some cardiac data looked like for individuals who lived at least another hour) and designing animal experiments.

7) **Presentation of results in an open forum.**
Systems like LSTAT are creating new data. The combination of the data from the multiple sensors that will be informatively combined in Activity 1 is effectively a new medical measurement. New data lead to new science. An important activity will be to present the information created in Activity 1 to medical and physiological researchers so that we may leverage the knowledge in the research community to improve the diagnoses made based on the new data being generated in this project.

G.4 Timeline by Task

1) **Detailed analysis and presentation of a patient's condition**
   a) Get data from the sponsor: October 1 – November 11
   b) Algorithms and code to obtain data features: October 1 – June 1.
   c) Dimension reduction: January 1 – September 1
   d) White paper on features: August 1 – September 30

2) **Visualization of the patient's current state**
   a) Candidate single patient views and interactions: January 1 – March 1
   b) Create prototypes of views and interaction: February 1 – July 1
   c) Present prototype and interaction: sometime during July 1 – September 30

3) **Cross-patient compressed views**
   a) Candidate cross-patient views and interactions: February 1 – April 1
   b) Create mock-ups of views and interactions: March 1 – August 1
   c) Present mock-up of views and interactions: sometime during July 1 – September 30

4) **Architectural issues for the MediSense system**
   a) Consult with client regarding eventual system: October 1 – December 15
   b) Architect: November 1 – April 1
   c) Presentation to client of issues and broad features of architecture.

5) **Handling dropped sensors or other missing/corrupt data**
   a) Determine state of art in handling missing data in decision analysis and clustering.
   b) Develop algorithms and test on real feature data.
   c) Presentation to the client on general strategies for handling dropped sensors, changes in sensors, and sporadically missing data.

6) **Patient states and sensor features**
   a) Consult with client regarding direction of investigation between patient states and sensor features: October 1 – November 15
   b) Gather information available from current data: November 1 – February 1
   c) Develop strategy for obtaining relationship between patient states and sensor features: January 1 – July 1
   d) Presentation of information and strategy: After July 1.

7) **Presentation of results in open forum**
   a) Sometime during fiscal year.

G.5 Deliverables

1) **Detailed analysis and presentation of a patient's condition**
   a) Presentations/demos to client
   b) White paper to client

2) **Visualization of the patient's current state**
   a) Presentation/demo to client of candidate views based on features estimated from Activity 1

3) **Cross-patient compressed views**

a) Presentation/demo to client of candidate cross-patient views and interactions with the views, along with discussions of relative merits of the candidates.

4) **Architectural issues for the MediSense platform**
   a) White paper outlining hardware and software issues for the MediSense platform, along with a plan for how MediSense might be constructed

5) **Handling dropped sensors or other missing/corrupt data**
   a) Presentation to client of issues, state of the art and strategy

6) **Patient states and sensor features**
   a) Presentation that describes some simple relationships, discusses the extent of the need for such information, and presents a plan for how the needed information will be obtained.

7) **Presentation of results in open forum**
   a) Submission of a conference paper or some other type of presentation at a professional conference. Includes preparing the article and participation in the conference.

G.6 Estimated cost

Tasks	Estimate
Detailed analysis and presentation of patient's condition	$95K-$110K
Visualization of patient's current state	$45K-$55K
Cross-patient compressed views	$40K-$50K
Architectural issues for the MediSense system	$35K-$45K
Handling dropped sensors or other missing/corrupt data	$30K-$40K
Patient states and sensor features	$40K-$50K
Presentation of results in open forum	$20K-$25K
Total	$305K-$375K

G.7 Leverage with Past PNNL Work

This work is strongly tied to the SPCE Task that is part of the Information Analysis and Visualization Research and Development Project for MRMC, client Major Stephen Bruttig. It is the natural outgrowth of the Digital Concept Exploration Task Sub-Task A: Researching options for managing and correlating information recorded from multiple sensors from multiple patients and Sub-Task B: Researching the feasibility of, and the options for, using the database created in A as a tool for recognition of symptoms.

In the SPCE Task, we have worked with one to three hours of electrocardiogram (EKG), A-line, and Blood Pulse Oximeter 1000 Hz sensor data from three patients recovering from open-heart surgery. We have developed a pattern-matching algorithm that picks the time of each heartbeat in the three related (but lagged) time-series and identifies the time period corresponding to a heartbeat. We have calculated simple

features and more sophisticated features based on the structure of the three sensor series for a particular heartbeat. We have also calculated cross-heartbeat features. These features are then used in an unsupervised clustering analysis to form groups based on feature similarity.

G.8 Principal Investigators

Paul Whitney
Paul.Whitney@pnl.gov
voice: (509) 375-6737
fax: (509) 375-2604

Nancy E. Miller
Nancy.Miller@pnl.gov
voice: (509) 375-6979
fax: (509) 375-2604

APPENDIX D

LANGUAGE TO MATHEMATICS

Language to Mathematics

Client: U.S. Army Medical Research and Materiel Command

December 23, 1997

Paul Whitney, Pacific Northwest National Laboratory

Contents

Figures

Tables

Introduction

This report describes research in language processing, the mathematical structure of language, and machine translation, performed by Battelle for the US Army Medical Research and Materiel Command. A key aspect of this work is to suggest areas for future research in language processing and translation. Reported in separate (attached) documents are:

1. Experimental results in creating mappings between documents in different languages, and
2. A web page containing various language processing references

This report begins with an overview of language processing, describing the typical types of problem it attacks. The references contained in the web page are particularly useful in offering an expanded view of language processing capabilities. The overview includes an introduction to mathematical representations of language. This is followed by a description of standard machine translation technology, based on a UCLA short course (Hovy and Knight 1997). To conclude the overview, we present a functional perspective that we find useful for organizing the vast language processing literature (which we refer to as the Language Processing Universe).

Next, some fundamental and regular mathematical properties of language are reviewed. We discuss how these properties have been exploited to address problems in language processing. Finally, we propose a language processing research agenda and some particular experiments which can further that agenda. Preliminary results from one of those experiments (in probability models for language) are presented in this report. The attached document on creating mappings between documents in different languages is also a beginning for one of the proposed experiments.

The proposed research agenda focuses in the near term on automatic assistance in obtaining the data needed for current designs of machine translation software. Examples of such data are "dictionaries" and grammars for languages. Currently, these are constructed laboriously "by hand". In the longer term, the research agenda focuses on estimating more abstract concepts (such as conflicts, engagements) from the data.

Overview of Language Processing

A quick overview of some applications of language processing is presented. Then a few of these topics are examined in more detail: information retrieval, summarization, disambiguation, and translation. Next, some mathematical and statistical regularities of language are discussed Finally, an organizing principle called the Language Processing Universe is proposed.

Applications of language processing

Common problems in language processing are listed below. We subsequently describe each problem in more detail.

- Document retrieval
  - Incremental retrieval
  - Query generalization
- Document routing
- Document browsing
- Document grouping
- Summarizing
- Document compression
- Speech recognition
- Editorial assistance
- Using context
- Language Identification
- Author Identification
- Translation
  - Dissemination
  - Assimilation
  - Parsing
  - Disambiguation
  - Text generation

Document Retrieval and Routing

Document retrieval is similar to gathering references for a focused question. Common examples of retrieval include using World Wide Web search engines to find information on particular subjects, or using a library's card catalogue. A key element here is that *one knows fairly precisely what the retrieved documents should be concerned with (at least in part).* Document routing is a related problem: given a document, which basket should one pitch it into or to which individuals and organizations should the document go? Retrieval and routing are similar, except that in retrieval there are only two classes: things you want to see and the rest, while routing will almost always involve multiple categories.

Incremental retrieval refers to a common aspect of how we typically do book research: we formulate initial queries, digest some of the information, and then formulate new queries based on the new information.

Query generalization refers to attempts to automatically create generalizations of a user's precise queries. Conceivably, a query generalization tool would reduce the number of iterations needed to get to the same amount of useful information.

Document Browsing and Grouping

Document browsing is almost an inverse of document retrieval: given a large collection of documents, what are they about (as opposed to knowing the latter and finding documents that satisfy the criteria)? The SPIRE package is an example of a document browsing system. Document grouping, dividing a large collection of documents into similar groups, is related to browsing: both activities are exploratory activities with "fuzzy" criteria.

Document Summarizing

Summarizing is creating automatic synopses of a document or collection of documents. The 1995 SIGIR had a nice example of a news story summarizer. The work appears to be very specialized and successful.

Document Compression

While not often considered as a major language-processing problem, compression of information is increasingly important. Papers in this area are published in language processing conferences (e.g. SIGIR 1995).

Speech Recognition

Converting human speech input into a form that generates the correct (that is, anticipated) response. There are commercial products available. We didn't look into the quality or utility of these products.

Editorial Assistance

Spelling and grammar checkers are examples. Another example of a useful tool would be an "assistant" that would turn an outline of a document, along with some research notes, into a rough draft of a research paper.

Using Context

Context is what enables you to correctly interpret a phrase like "Wow that's fast". It's a particularly difficult area in language processing.

Language Identification

Given a passage or recording, what language is it in? An interesting attack on this problem (for text) is given in Damashek 1994. It turns out that a useful discriminator among languages is the relative frequencies of n-grams.

Author Identification

An early example of the application of statistics to language processing was the identification of authorship of some of the Federalist papers, see Mosteller and Wallace (1984).

Translation

Translation refers to the conversion from one human language to another. There are commercially available products (see the reference web page) of useful quality. *Machine translation* refers to a computer program that performs the translation.

Dissemination and *Assimilation* refer to the intended use of translated language. Dissemination refers to translating with the intent of distributing documents in the target languages. Assimilation refers to translating with the intent of understanding the translated material. The distinction between these two applications is made because to be useful, dissemination translation generally needs to be of higher quality than assimilation translation.

Parsing refers to the step in a machine translation system that's analogous to diagramming a sentence (i.e., analyzing parts of speech and the sentence structure).

Disambiguation is the problem of deciding which sense of a word (or common phrase) is meant. For instance, the word "plant" has both biological and industrial meanings or connotations.

Text generation is that part of translation (and some other language processing activities, such as text summarization) which refers to generating fluent text from an abstract representation. For instance, given a diagram of a sentence, a text generation problem is to write the actual sentence.

Vector Space Model in Information Retrieval and Browsing

The current state of the art in information retrieval is to map documents into a vector representation in such a way that:

1) each document corresponds to a single vector,

2) similar documents are represented by vectors that are close together, and

3) dissimilar documents are represented by vectors that are far apart.

A crude vector representation for a document is a listing of the words that appear in the document along with their frequencies of appearance. Numerous refinements to this document vector are possible. For instance, frequently used conjunctions, articles, and similar words (such as "and", "the", "a") are often not included in the summary. The rationalization for excluding these words is that they do not directly reflect the content of the document. The basic idea goes all the way back (at least) to Salton (1971); see Faloutas and Ord (1995) for a recent survey.

Another word frequency-based vectorization is the collection of vectors that corresponds with the collection of documents. The collection of vectors is represented as a matrix; the entry in the matrix for the i^{th} row and j^{th} column is the word frequency for the i^{th} word and the j^{th} document. This matrix is then approximated by a smaller matrix with fewer rows; the approximation being constructed using the singular value decomposition. The resulting vectorization is known in the information retrieval community as latent semantic indexing; see the Bellcore web site (http://superbook.bellcore.com/~std/LSI.html) for numerous references and applications.

For document browsing, the vectors can be presented in a variety of ways. The SPIRE system projects the high-dimensional vectors into a viewable 2 or 2.5-dimensional space; see Wise et al (1995). Another visualization strategy for browsing is to project the document vectors onto a grid via self-organizing maps (Kohonen 1995). This strategy is discussed in Honkela et al (1996).

Case Study: Text Summarization

An experiment in summarizing news articles about terrorist events is documented in McKeown and Radev (1995). The input data are the news articles related to a particular terrorist event; the output summary is a fluent, English synopsis of the articles. The synopsis also includes summary information about how the story was reported over time.

The entire system, as reported in the cited paper, is not completely automatic; the stories containing the thread of the description of an event are selected by hand. Information from each story is extracted and placed in a template; and some of the templates were filled in by hand. Based on the collection of templates, the summary is constructed by the system.

Mathematical Representations of Language

We discuss three fundamental aspects of language that have strong links to verifiable regularities in language. The first is general frequency or probabilistic properties of language. This work goes pretty far back (to the 30s for Zipf's contribution; Shannon's contributions began to appear in the 1940s). We propose new classes of models to be considered in this same context. Initial work suggests that these models are worth pursuing. The second sub-section describes work in unsupervised "part of speech" tagging. An intriguing aspect of this work is the use of commonly occurring words as "stakes in the ground" or points of reference from which one can infer grammar and parts-of-speech. The third sub-section describes an algorithm that can be used to disambiguate words.

Interestingly, both part-of-speech tagging and disambiguation have different objectives and yet apply similar mathematical techniques (such as cluster analysis or something very close to it) that are based on similarity of certain features (other words in the neighborhood).

Frequency/probabilistic properties of language

A fundamental observation about language is that the occurrence of words follows a pattern. Ordering the words in a large collection of text from most frequent to least frequent, and letting w_i denote the number of occurrences of the i^{th} most frequent word, Zipf's law states that:

$$w_i \propto \frac{1}{i}$$

A curiosity of Zipf's law is that it's not a probability distribution (not for an infinite number of words, anyway). The relation could be used as a constraint on candidate language models.

More detailed models of language are based on estimating the probability that a word will occur next, given some number of the immediately preceding words; for instance

$$x_i | x_{i-1} x_{i-2}$$

represents the probability that the word x_i occurs given the two immediately preceding words were x_{i-1} and x_{i-2}. To make this more concrete, consider the two words

$$\text{open the} \begin{cases} \text{is} \\ \text{gestured} \\ \text{door} \\ \text{verbally} \end{cases}$$

It's natural to guess that the next word to occur is more likely to be "door" than "is", "gestured" or "verbally". These model types are formally described as Markov models (the model in the example above is "second order" since it looks back two words). The data needed to estimate the model are the frequencies of occurrence of word triplets and word pairs. These word combinations are referred to as "n-grams" of words. So a bi-gram (based on the previous sentence) is "These word"; a tri-gram in the previous sentence is "word combinations are" and a quad-gram is "These word combinations are".

In the context of Markov models for language, Zipf's law is a zero-th order model. Second and third order models have been used as part of text generation (discussed in a subsequent section of this paper) and machine translation, see Brown et al (1992).

Zipf's law provides a basic expectation that the number of unique words will grow "slowly" relative to the number of words in a document collection. Figure 1 provides some basic data on the frequencies of unique n-grams of words. The figure shows that the typical tri- and quad-gram is never repeated in the document. The large amount of text needed to estimate even the second-order Markov model is a barrier.

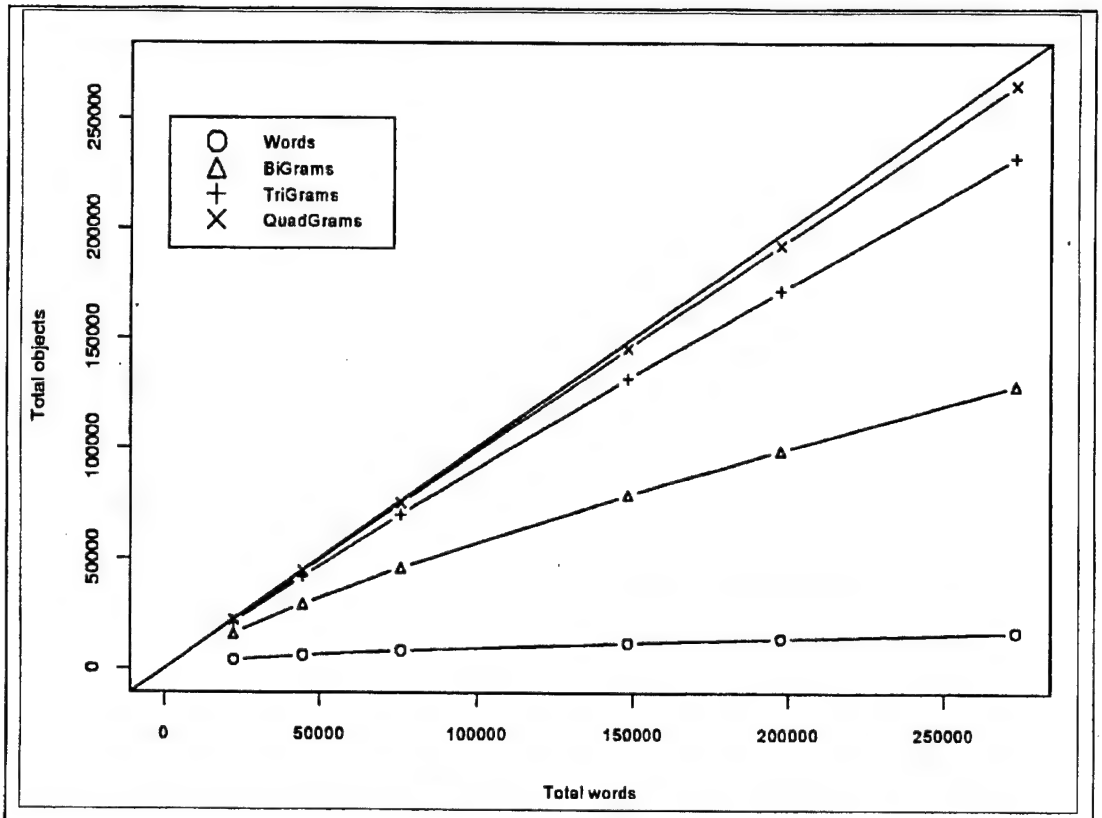


Figure 1: Frequency of the unique words, bi-grams, tri-grams and quad-grams as the number of words increases.

Markov models have also been applied to text at the character level. These ideas go back (at least) to Shannon (1950); see Cover (1991) for an exposition of this work. Recent applications of character based probability models for text are given in Damashek (1994, 1996). In this work, frequencies of occurrence of "n-grams" of characters are used as features to characterize and distinguish text. Damashek presents application of character n-grams to the problem of language identification.

Alternative Probability Structures for Language

Application of the simple word-based n-gram models described above is problematic: often there's not enough data, *and* model fit becomes increasingly questionable as larger sequences are considered. However, probability models apply directly to various language processing problems and algorithms; hence, any progress or improvement in these probability models translates into progress in multiple LP problems.

Two simple possibilities for improved probability models for text are given below. The first is based on a standard projection idea. In the second, one might imagine "neighbors" being constructed based on grammar or proximity to certain key words:

$$\frac{1}{2}x_i \mid x_{i-1} + \frac{1}{4}x_i \mid x_{i-2} + \frac{1}{4}x_i \mid x_{i-3}$$

or

$$x_i \mid \text{neighbors}(x_i)$$

The first proposed model is intended more as an instance of a general class than as a firm

suggestion. The general class intended is mixtures of models that condition on only one component at a time; these models are discussed in Elliot et al (1995) and MacDonald et al (1997). In related curve fitting problems, this type of alternative model has led to significant progress by reducing the effective dimension of the variable being conditioned on. In the case above, we're considering four variables in a row (up to N^4 possibilities in a full quad-gram model), whereas with the proposed model we only have to keep track of at most $3*N^2$ potential pairs (assuming a vocabulary of N words).

The second proposed model is intended to suggest that we consider different types of neighbors (perhaps using a "syntax" based on distance, e.g., nearest verb or adverb) to construct our conditioning. Much thought still needs to be given to the form of the neighbor function(s).

The unanswered question here is whether these proposed models fit the data better than bi- or tri-gram models. Tri-gram models are the ones to beat for accuracy. Bi-gram models are used because they're easier on computer resources than tri-gram models and sometimes accurate enough.

Both of these proposals are original; although they're "natural" enough that we will not be surprised if additional investigation turns up someone who is working on them.

A "goodness of fit" criteria used in Brown et al (1992) to compare the fit of different probability models on text is called the *perplexity.* The perplexity of text "S" is

$$\Pr(S)^{-1/|S|}$$

where the probability $\Pr(S)$ depends on the model (for instance, the order of the Markov process) and $|S|$ is the length of the text (in words). The natural log of the perplexity can be interpreted in the context of mathematical statistics as the average log-likelihood; see Lindgren (1976). Higher values indicate a better agreement between the model and the data. In Brown et al (1992), the perplexity was estimated by using part of the available text to determine the model parameters, and using the remaining text to estimate the perplexity.

A recurring problem is how to estimate the contribution to $\Pr(S)$ from n-grams of words in the evaluation data set that did not occur in the training data set. Table 1 below illustrates this difficulty.

Table 1: Number of missing n-grams from evaluation data.

			Proportion Missed			
Model Text	Evaluation Text	Number of words in model text	1-grams	bi-grams	tr-grams	quad-grams
Books 1,2	Book 8	44759	0.09	0.53	0.89	0.98
Books 1,2	Book 9	22357	0.09	0.53	0.89	0.98
Books 1,2,3,4	Book 8	44759	0.07	0.39	0.85	0.97
Books 1,2,3,4	Book 9	22357	0.07	0.38	0.88	0.97

Table 1 the proportion of n-grams (n=1,2,3,4) that do not occur in the evaluation text for given mode texts. The proportion of missing n-grams is high for every "n" except 1. For tri- and quad-grams, the majority of n-grams encountered in the evaluation text are missing. The table points out the need for large data sets and smart models that take advantage of linguistic knowledge.

Unsupervised Syntactic Word Tagging

A fundamental property of language is that words are created and destroyed at a faster rate in certain syntactic classes than in others. A closed class of words is a part of speech (that is, a collection of words) to which new words are not added. For instance, prepositions form a closed class; nouns are an open class (new nouns are created all the time). It turns out that most words

that occur frequently are closed class words.

It's been additionally observed that closed class words serve as (part of the) framework for language. A formal statement of this idea is the *replacement test*:

Replacement Test: If a word or phrase has the same distribution as a word or phrase of a known type; then it is a word or phrase of that type

e.g., *red* and *broken* are often of the same syntactic type:

> The *red* light is on.
>
> The *broken* light is on.

While the second sentence is nonsense, it's also grammatically correct. *Red* and *the color of blood* are often not equivalent:

> The color of blood light is on.

The replacement test hinges somewhat on subjective judgments about similarities in the "goodness" of sentences; however, it's an interesting guideline, which has been used to categorize words. The implementation of the replacement test is consistent with there being "degrees of similarity". The context also affects this judgment. The relative position of a word with respect to the most frequent words is strongly related to "parts of speech" information. Finch (1994), among others, ran an experiment whose results support this claim.

The experiment reported in Finch (1994) proceeded by first identifying the 150 most frequent words in the test data (news text). Next, a vector for each word in the text was constructed by noting the relative position of the word with respect to these 150 frequent words, in a neighborhood up to 5 words in size; an entry of 1 indicates the presence of a particular frequent word, 0 its absence. These vectors for each unique word were then used to create a vector upon which a clustering of words was based. Figure 2 shows some of the clusters resulting from this experiment. It can be seen that the clusters contain common parts of speech; and that for some of the clusters, the meanings of the words are similar. However, the similarity in meaning does not hold up across clusters.

C49 work sound end act
C50 post report show answer fix quote match sign
C51 so
C52 right wrong fine ok
C53 seems appears feels says thinks knows knew tells asks
C54 two three four five six ten seven
C55 sure guess bet suppose suspect doubt
C56 well far
C57 fact course least
C58 working running playing moving reading writing driving flying walking
C59 experience sense advantage effect attention interest concern
C60 evidence proof truth meaning reality nature existence belief success loss religion faith
C61 long fast early late cold
C62 interested concerned aware involved responsible familiar
C63 years months weeks days hours minutes
C64 help care
C65 here today
C66 science engineering management physics art computing programming training processing communications communication
C67 israel iraq india america china japan kuwait europe canada taiwan
C68 set cut hit beat shut
C69 now
C70 myself yourself themselves itself himself
C71 able willing ready
C72 x ftp uucp nfs

Figure 2: Word clusters from neighborhood vectorization; from Finch (1995).

The methodology described by Finch is entirely language independent.

Disambiguation

Disambiguation is a technical term in language processing that means to decide which meaning of a word is intended. Yarowsky (1995) presents a data driven methodology for disambiguation. While data driven, the methodology also incorporates some fundamental observations about language, viz.:

1. Nearby words (context) determine the sense of the target word
2. The sense of a word is very consistent within a given document

e.g., the word *life* occurring within 10 words of the word *plant.*

Yarowsky does an empirical check on the second principle, and it appears to hold up. One could weaken the principle somewhat to "The sense of a word is very consistent within a neighborhood extending a few paragraphs from the target word", and perhaps still get useful disambiguation information.

Table 3 below is from Yarowsky (1995). The table shows various instances of the word plant along with some of the context. A human reader is NOT hard pressed to distinguish two senses of the word plant from this table.

Table 2: The word *plant* in contexts

Sense	Training Examples (Keyword in Context)		
?	... company said the	*plant*	is still operating
?	Although thousands of	*plant*	and animal species
?	... zonal distribution of	*plant*	life . ...
?	... to strain microscopic	*plant*	life from the ...
?	Vinyl chloride monomer	*plant*	, which is ...
?	And Golgi apparatus of	*plant*	and animal cells
?	... computer disk drive	*plant*	located in ...
?	... divide life into	*plant*	and animal kingdom
?	... close-up studies of	*plant*	life and natural
?	... Nissan car and truck	*plant*	in Japan is ...
?	... keep a manufacturing	*plant*	profitable without
?	... molecules found in	*plant*	and animal tissue
?	... union responses to	*plant*	closures . ...
?	... animal rather than	*plant*	tissues can be
?	... many dangers to	*plant*	and animal life
?	Company manufacturing	*plant*	is in Orlando ...
?	... growth of aquatic	*plant*	life in water ...
?	Automated manufacturing	*plant*	in Fremont ,
?	... Animal and	*plant*	life are delicately
?	Discovered at a St. Louis	*plant*	manufacturing
?	Computer manufacturing	*plant*	and adjacent ...
?	... the proliferation of	*plant*	and animal life
?	...		...

The implementation of disambiguation proceeds as follows:

- Isolate a small number of training examples representative of various word senses
- Select rules (of the form "animal within a 5-word neighborhood"), based on how well they distinguish word senses in the training sets
- Expand the training set using rules that perform well, and retest.

The methodology used is strikingly similar to cluster analysis, but is not exactly cluster analysis.

Text Generation

Text generation is the problem of generating fluent text to correspond with an abstract representation. Text generation is used in automatic summarization and machine translation. In machine translation, text generation can be accomplished in many forms. Machine translation can lead to numerous candidate output sentences. Knight and Gatzuvassukigkou (1995) show how simple probability models for language can be used to decide among the options efficiently. Figure 3 shows an example from Knight et al. (1995). A random set of selections from the candidate translations is shown, as well as the highest-ranking translations, where the rank is calculated from the probability of the sentence. As the figure shows, the probability-based ranking works well.

RANDOM EXTRACTION

- Her incriminates for him to thieve an automobiles.
- She am accusing for him to steal autos.
- She impeach that him thieve that there was the auto.

STATISTICAL BIGRAM EXTRACTION

- She charged that he stole the car.
- She charged that he stole the cars.
- She charged that he stole cars.
- She charged that he stole car.
- She charges that he stole the car.

Figure 3: Example of application of probability models in text generation; from Knight et al (1995)

Multiple candidate sentences are available from machine translation (MT) systems after the syntax has been mapped to the target language and candidate word replacements have been created. How does one choose among all these candidates?

If a probability model is available for the target language, then one can choose the most likely sentence (or from among the likely sentences). Figure 3 shows some candidate sentences, and the likely ones. The likely sentences are fluent; the typical sentences are very clumsy.

Work in this field points out an interesting research opportunity: tuning the probability model to produce a certain style/dialect of the language. For instance, one might want to "tweak" the underlying probabilities to slightly increase the probability of various dialects or writing styles.

Machine Translation

Machine translation is a growing part of the translation industry. Estimates put the total volume of translation business at $20 billion in 1989 (Hovy and Knight 1997). This section describes the current technology in machine translation, and points out opportunities for research.

MT is available at the personal computing level. For instance, numerous companies sell programs to translate between common languages, and CompuServe© has translation services embedded in its email and chat room services.

Parts of MT technology that are working well include name recognition, parts of speech tagging, morphology (working with tense and other transformations of a word root). Parsing works well, although much parsing technology is handcrafted. There are on-line lexicons that are used in machine translation (again, the construction of many of these was labor intensive).

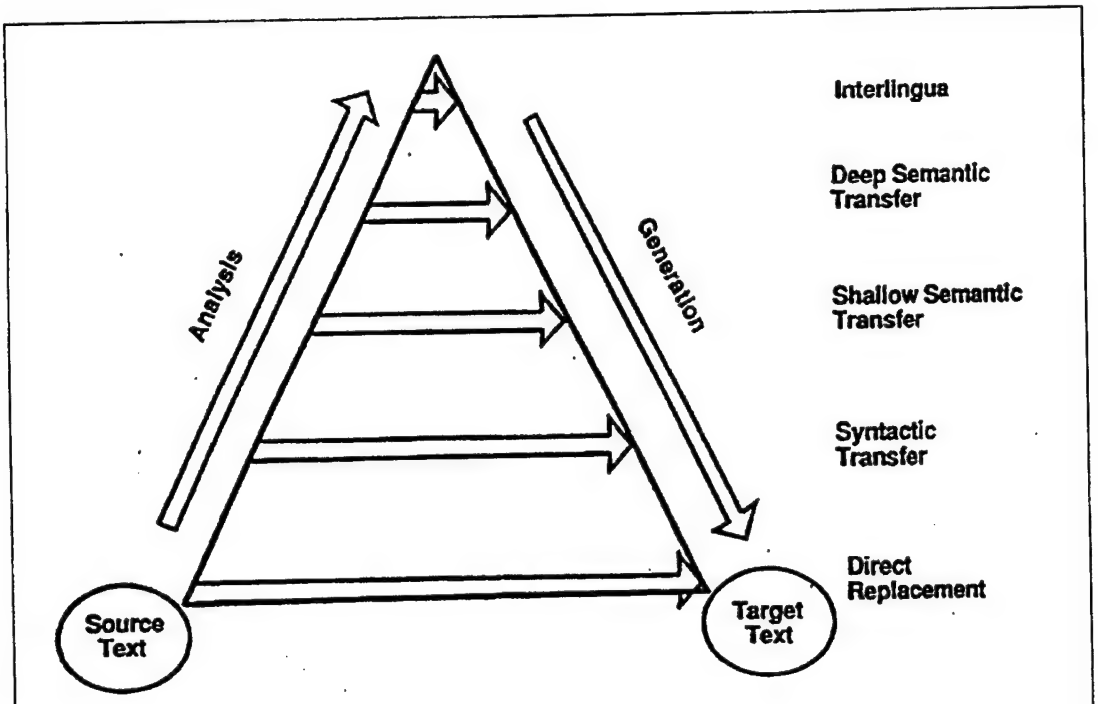


Figure 4: Various options for machine translation architectures; this slide is from Ed Hovy's course notes on Machine Translation[1].

The triangle in Figure 4 represents possible architectures for an MT system; or perhaps better said, the triangle represents architectures currently used for MT systems. The interlingua component is like the "Objects, Relations among Objects" component of the LP Universe model, discussed in the following section.

The triangle "works" as follows: different types of translation are represented by moving up from the Source text, across, and then down to the Target text. For instance:

"Direct replacement" translation is done by replacing the words in the source text with their dictionary equivalents in the target language. This type of translation corresponds with moving directly across from the Source text to the Target text.

"Syntactic transfer" translation is done by parsing the input source sentence. Then the parsed sentence is mapped to a parse (or perhaps multiple possible parses) in the target language. Finally, the target language words are filled into the structure and the full target text is generated.

"Semantic transfer" translation refers to determining the correct interpretation of words. As indicated, there are various degrees or levels of effort available, ranging from shallow semantic analysis to interlingual analysis.

"Interlingua" refers to an abstract, language-neutral, description of the world. Thus, mapping a language into an interlingua amounts to mapping the language into another whole (possibly new) language. The Kant project at Carnegie Mellon University is constructing an interlingual translator for the technical manuals of the Caterpillar Corporation. A potential benefit of interlingual translation over "all possible language pair" translation is the saving in effort associated with having to construct only translators between each language and the interlingua (a total of N

[1] See http://www.isi.edu/natural-language/people/hovy.html

translations for N languages) versus having to construct translators between each possible pair of languages (a total of N(N-1)/2 translators for N languages).

Problems in MT

At the word level, the following are common challenges:

- Morphology: finding root forms is a diverse challenge, depending on the language.
- Word isolation: many languages are written without interword spaces.
- Spelling errors and segmentation errors: both happen all the time.
- Ambiguity: resolving the meaning of a word
- Transliteration of foreign names: For instance, note the numerous versions of "Peking" in English text.
- Multiple reference forms: "Mr. President", "Bill Clinton", "he" can all refer to the same person.
- Metaphors – a problem at the phrase level, e.g., "The bears are loose on Wall Street".

At the sentence level, common challenges are:

- Parsing – ungrammatical input is possible and often "correct" (e.g., in slide presentations and conversations).
- Structural ambiguity – can be due to missing words, dangling participles, etc.
- Semantic ambiguity – can be due either to missing information or missing context; for instance: "Bill forgot his anniversary. His rent is due tomorrow."

At the multi-sentence level, there are problems related to the structure of a document or a conversation. For instance, the typical form of a quote is a structure that people commonly encounter and handle correctly. There are numerous other special case document structures as well: presentations, technical reports, business letters, forms etc. Handling these in the context of machine translation is an outstanding challenge.

Current Challenges and Research in MT

Currently, most work in MT systems is involved in creating the lexicon. The information desired in a lexicon includes spelling forms, syntactic information, and semantic information. The approaches used to construct a lexicon are 1) manual listing 2) mining of dictionaries and 3) mining of text. The latter two are preferable from a cost perspective.

Generally, the problem of constructing a lexicon is a problem of knowledge acquisition; this problem is common to MT and most artificial intelligence applications. Recent efforts focus on automating the acquisition of the needed knowledge. For MT, the possible sources include raw text (plentiful in many languages), bilingual text (less plentiful, but still available), online monolingual dictionaries, online bilingual dictionaries, online thesauri, etc. The data are extant; the need is for algorithms and experience.

The previous sections on automatic parts of speech tagging and disambiguation provide strong clues about how to construct a lexicon from raw text. Pursuing this line of research is a highly recommended activity.

Language Processing Universe

A caricature that can be used to compare various work and strategies in language processing and translation is the language-processing universe (LPU). The LPU is diagrammed in Figure 5. The LPU shows relations between language, the structure of language, and "reality" (or at least some abstract model of reality). Structures of language include grammar, various mathematical

representations, and models (such as the probability and vector space models described earlier). The diagram suggests that language, as a mathematical object, is constrained by the mappings shown in the diagram. The rest of this section shows how several of the examples and activities described above map into the LPU.

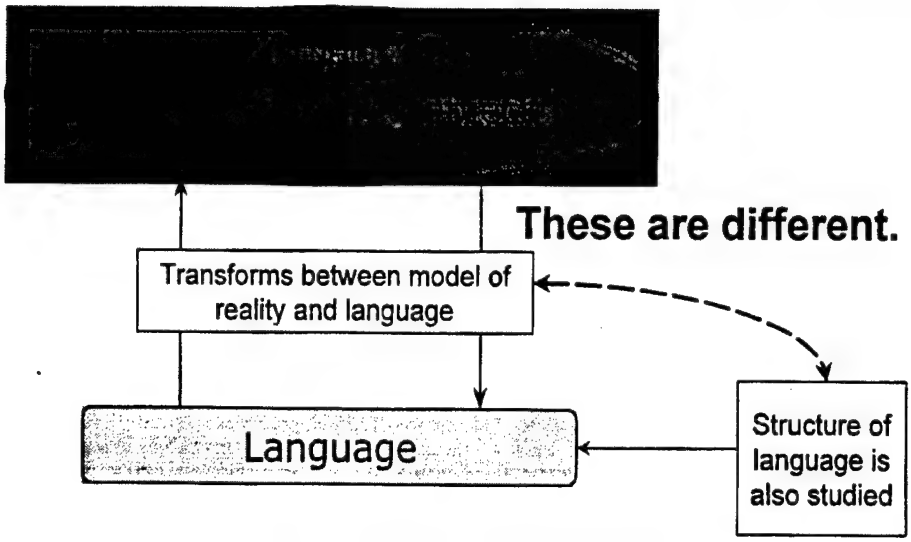


Figure 5: Language Processing Universe

Figure 6 shows the parts of the LPU used in probabilistic models of text. Regardless of whether the models are n-grams of words, n-grams of characters, or include information about syntax or parts of speech, all of the "structure" is intrinsic to the language or text part of the LPU. The LPU for the vector space model would be similar; and the language structure would be the vector representation, see Figure 7.

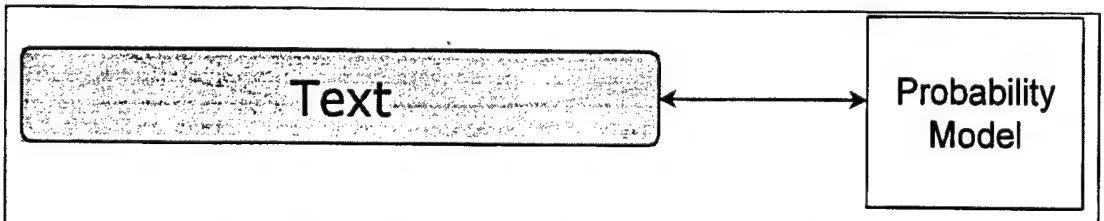


Figure 6: The LPU as applied to probability modeling of text.

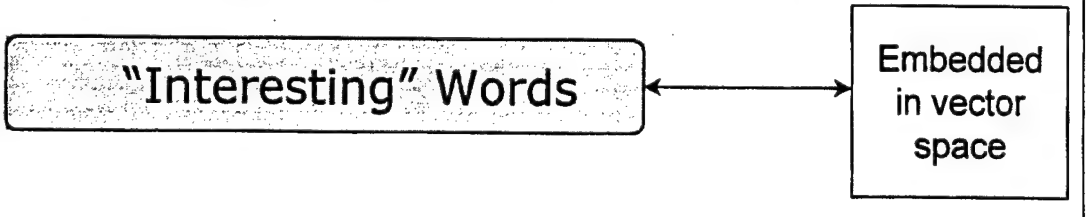


Figure 7: The LPU for the vector space model used in information retrieval

The summarization application described earlier exercises all of the LPU if the following identifications are made:

- The templates are the "real world" information;
- The "objects" are incidents, people (victims), and news organizations;

- The states are the type of incident, the number of people killed, the date, and which news organization is reporting. Once the information from a news report has been gleaned and stuffed into the template, it becomes a challenging but possible task to create a "nice" news summary for this type of event.

In terms of the LPU, a very narrow set of reality and language is combined, and grammar is applied to produce a nicely formed summary. The LPU is the complete package for a small application; see Figure 8.

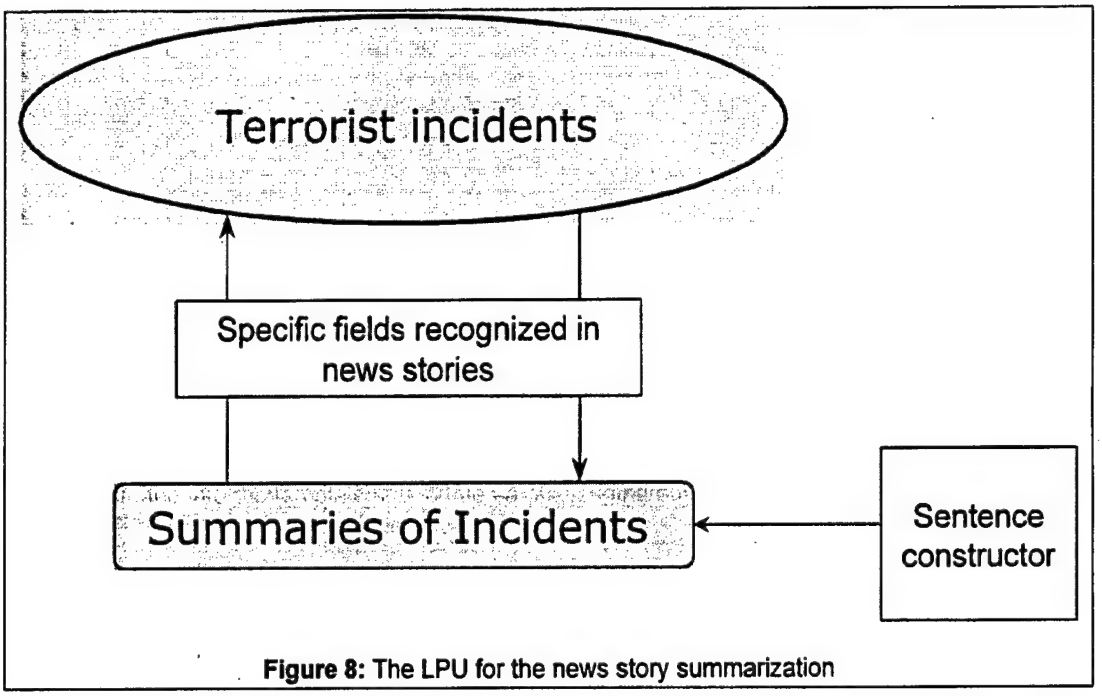


Figure 8: The LPU for the news story summarization

Variations on this diagram are found in the literature. For instance, the translation triangle shown in Figure 4 can be mapped into the LPU (Figure 5) if one identifies "reality" with the Interlingua. The complete LPU is typically not fully utilized, except in focused application areas. The following observation seems to hold: For sufficiently focused applications, the LPU can be completely utilized (see Devlin 1997 for additional discussion of this observation).

Research Agenda and Objectives

This section outlines a research agenda for machine translation and language processing. The agenda is divided into short term, intermediate term, and long term activities. The agenda items called out under the first two categories can begin now (and in some cases have already begun). Several candidate research experiments are proposed and discussed in light of this agenda.

Short Term MT/LP Research

- Alternative probability structures for language
- Updating probability structures for representations of language
- Automating acquisition of lexicon and other components of MT systems

The above research areas were gleaned from the literature, related PNNL projects, and Hovy and Knight (1997). Indeed, numerous possibilities exist for near-term research; the list above reflects current PNNL biases.

Intermediate Term MT/LP Research

- Common features across languages
- Matching text to known scenarios
- Application of visualization in MT

There is a growing body of work in the area of matching text to scenarios, as this matching is a variation on information retrieval problems that are currently under evaluation in the TREC program[2].

A conjecture about how automated learning (or at least natural language understanding) could evolve is as follows:

1) Learn how to glean the "things" (nouns) from text or other information sources.

2) Learn how to put the "things" together based on their interactions (verbs) and states (adjectives, other descriptors).

3) Take the information from 2), and map it into scripts or scenarios.

While there are tools for both 1) and 2), there is significant appeal in working with language-independent strategies. Work done to map text into scenarios describing eating at a restaurant is documented in Miikkulainen (1993); this book summarizes work related to understanding text descriptions of focused situations. However, the work described there is based on limited text samples. Learning how to map text into a more extensive collection of scenarios/scripts is a natural extension of previous work; and might naturally lead into "machine learning".

One of the most useful developments would be to build (visualization) tools that can be used to examine text; and provide these tools to the linguistics community for critique and use. Virtually no work is being done in this area.

Long Term MT/LP Research

- Inferring scenarios from examples
- Automated learning in a specific field (e.g., medicine) from documents and other media
- Interlingual translation of medical documents

---

[2] See http://trec.nist.gov/overview.html

- Good MT from examples

Comment: These are more objectives than specific research agenda items, and are probably best used to evaluate how "on track" the intermediate term and short term research is progressing.

Candidate Experiments

This section describes candidate experiments in language processing and machine translation technologies. A brief description of each proposed experiment is provided, along with a list of the pros and cons of undertaking that experiment. Work toward two of these experiments is underway: preliminary work in alternative probability models was described earlier in this document; and work in understanding how to use multilingual text data to create mappings between documents in different languages is described in Whitney and McQuerry (1997).

Experiment: Parallel corpus language analysis

Fundamental to understanding language in general and translation in particular is the analysis of documents in diverse languages. A parallel corpus is a collection of documents that are available in multiple languages. An experiment conducted this year on a parallel corpus shows that a strong mapping can be constructed between documents in different languages (Whitney and McQuerry 1997). We propose to take a further step and empirically estimate mappings between more detailed structures in the different language documents. The particular structures of interest include words, phrases, syntax, sentences, and morphology.

- Pro: Includes steps needed for "data based" machine translation
- Pro: Requires continued exposure to multi-language data
- Cons: none

Experiment: Alternative probability models

The proposed experiment is to examine the extent to which different probability models can be used to fit text. A useful outcome would be to obtain models that fit text well without requiring as much training data as current n-gram based models.

- Pro: If a better fitting model can be found, applications are numerous and "immediate"
- Pro: Opportunity to get into syntax
- Pro: Connected to long term R&D objectives
- Pro: Opportunity to consider how to visualize these probability structures
- Pro: Numerous collaboration opportunities
- Pro: Cross-cutting impact if successful, due to ubiquity of probability models in MT/LP.
- Cons: none

Experiment: Parts-of-speech tagging based on neighboring words

The proposed experiment is to build and test a data driven parts-of-speech tagger for English and some other language, utilizing the principles outlined in Finch (1995).

- Pro: Opportunity to get computationally deeper into language
- Pro: Good fit with intermediate and long term research agenda – this is a needed step in scenario/script understanding
- Pro: Properties upon which the tagger is based apply to many languages
- Pro: The generated information is useful for MT
- Con: Taggers now exist for English and other common languages and are freely available

Experiment: Word pair document vectors

The proposed experiment is to extend the current vectorization in SPIRE, to be based on successive word pairs, as opposed to single words.

- Pro: Expect some information to be provided about semantics
- Con: Information might not "collapse" to a computationally tractable size, although some pre-processing along the lines of hierarchical work might ameliorate this problem
- Con: Not well tied to the research agenda

The expectation regarding semantics is based on the disambiguation work of Yarowsky, which showed how context could be used to deduce the senses of a word.

Experiment: Parse trees and word frequencies

The experiment is to combine information based on the syntactic content of a document with word usage to create a new document vectorization.

- Pro: Opportunity to get into syntax
- Pro: Potential academic applications to Federalist papers and to authorship "disputes"
- Con: Uncertainty that syntax combined with word usage will yield meaningful vectors
- Con: Weak connection with R&D Agenda, except for construction of parse trees.

The word usage component of the information can be calculated using current software. The syntactic component will require 1) capability of parsing and 2) vectorizing a parse representation.

Experiment: Tweak probability models to generate dialects

The experiment would be to generate different dialects of text based on 1) an abstract representation of a collection of sentences and 2) adequate samples of text from the target dialects.

- Pro: Directed at current MT problem area
- Pro: Collaboration opportunities available
- Pro: Connected with intermediate term and long term agenda
- Con: Idea seems to have limited scope or impact in language processing

Conclusions and Recommended Activities

Parts of the proposed agenda can be conducted elsewhere (Universities, industry etc.) because the agenda items are relatively well defined and the anticipated return is fairly immediate. We've excluded those parts of the agenda and have isolated three parts of the agenda for PNNL to attack. They are:

- Alternative probability models for language
- Multi-lingual language processing
- Matching text to known scenarios

Alternative models are worth investigation by PNNL because of the uniquely diverse combination of skills available at PNNL: in particular the combination of language processing and mathematical modeling experience. Additionally, this work contributes to building expertise in MT and continues language experiments begun in FY97. Common features across languages (especially probabilistic and related structural features) are not widely studied. We anticipate that understanding these features will be necessary for MT and multi-language processing. Finally, the activity of inferring aspects of a particular scenario from text is important from the perspective of the long-term goal of inferring classes of scenarios from a collection of documents. This activity steps beyond the boundaries of the SPIRE technology.

The final recommended activity is to revise the research agenda for MT and LP. We view this activity as maintaining contact with the MT and LP communities, considering how various advances affect and change the R&D agenda described here, and updating it.

References

Brown, Peter F., Peter V. deSouza, Robert L. Merver, Vincent J. Della Pietra, Jennifer C. Lai. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics* **18**-4 pp467-479.

Carbonell, Jaime, Yiming Yang, Robert Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. "Translingual Information Retrieval: A Comparative Evaluation". *In Proceedings of Fifteenth International Joint Conference on Artificial Intelligence* (IJCAI-97).

Charniak, Eugene (1996) *Statistical Language Learning*. MIT Press.

Charniak, Eugene (1995) Parsing with context-free grammars and word statistics, Technical Report CS-95-28, Department of Computer Science, Brown University.

Damashek, Marc (1994). Gauging Similarity via n-grams: Text Sorting, Categorization and Retrieval in any Language. TR-R53-05-94, National Security Agency.

Damashek, Marc (1995). Gauging Similarity with n-Grams: Language-Independent Categorization of Text. Science 267 p844-848.

Devlin, Keith J. (1997). *Goodbye, Descartes : The End of Logic and the Search for a New Cosmology of the Mind.* John Wiley & Sons, New York.

Dumais, S: T., Landauer, T. K. and Littman, M. L. (1996) "Automatic cross-linguistic information retrieval using Latent Semantic Indexing." In SIGIR'96 - Workshop on Cross-Linguistic Information Retrieval, pp. 16-23, August 1996.
http://superbook.bellcore.com/~std/papers/SIGIR96.ps

Elliott, Robert J., Lakhdar Aggoun and John B. Moore, (1995). *Hidden Markov Models: Estimation and Control.* Springer-Verlag, New York.

Faloutsos, Christos and Douglas W. Oard. (1995). "A Survey of Information Retrieval and Filtering Methods." Technical Report CS-TR-3514. Dept. of Computer Science, Univ. of Maryland.

Honkela, Timo, Samuel Kaski, Krista Lagus, and Teuvo Kohonen, (1996). "Self-organizing maps of document collections", *ALMA* **1**-2. Electronic Journal, address is http://www.diemme.it/~luigi/alma.html

Hovy, Eduard, and Kevin Knight (1997). Machine Translation. Course notes from University of California, Department of Engineering, Information Systems and Technical Management Short Course Program. http://www.unex.ucla.edu/shortcourses/spring97/mach.htm

Kipf, G.K. (1935) *The Psychobiology of Language*. Houghton Mifflin, Boston.

Kohonen, Teuvo, 1995. Self-Organizing Maps. Springer-Verlag, Berlin.

Knight, K., and V. Hatzivassiloglou (1995). "Two-Level, Many-Paths Generation," *Proc. of the Conference of the Association for Computational Linguistics* (ACL).

Lindgren, Bernard W. (1976). Statistical Theory. MacMillan Publishing Co. New York.

MacDonald, Iain L. and Walter Zucchini (1997). *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman&Hall.

Mosteller, Frederick and David L. Wallace (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, Springer-Verlag, New York.

Samuelsson, Christer, (1996). Relating Turing's Formula and Zipf's Law. *In Proceedings of the 4th Workshop on Very Large Corpora*, pp 70-78, ACL. Also available as CLAUS Report 78; cmp-lg/9606013.

Salton, Gerard (1971). *The SMART Retrieval System – Experiments in Automatic Document*

Processing. Prentice-Hall Inc, New Jersey.

Turtle, Howard and W. Bruce Croft (1992). "A Comparison of Text Retrieval Models". *The Computer Journal*, **35**-3 pp279-290.

United Nations Parallel Text Corpus (English, French, Spanish) 1994. Available from the Linguistic Data Consortium,

http://www.ldc.upenn.edu/ldc/catalog/html/text_html/unptc.html.

Wise, J.A., Thomas, J.J. et al 1995. Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. *Proceedings of the IEEE '95 Information Visualization Conference*, 51-58.

Yarowsky, D., (1995). "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics,* pp. 189-196, Cambridge, MA.

APPENDIX E

APPLICATIONS OF PARALLEL CORPORA IN DOCUMENT ANALYSIS

Applications of Parallel Corpora in Document Analysis

Paul Whitney, Dennis McQuerry

Pacific Northwest National Laboratory

1. Introduction

An information retrieval issue that will become increasingly important is the preponderance of documents in multiple languages. Challenges associated with these data objects include inferring relations among the documents, retrieving them, and searching them. In this paper, we investigate the potential of the language analysis tool embedded in SPIRE (see [SPIRE95]) for addressing multi-language document collections. In particular, we estimate the relationships between English, French and Spanish versions of the same documents.

A parallel corpus of documents was obtained from the Linguistic Data Consortium (LDC94). Sufficient pre-processing was done to ensure that we had a parallel corpus, that is, the same documents in different languages. Then we processed the documents in two ways: one in which all the documents were aggregated into a single corpus, and another in which the overall corpus was divided into separate English, French, and Spanish collections. For each mode of processing, the SPIRE output included the vectorization of each document (each document was mapped into a high dimensional vector) and the clusters of documents that were calculated based on the vectors.

We viewed the documents in a combined space to see whether the SPIRE tools would correctly separate the documents according to language; see Figure 1 for the resulting projection. We also compared the separate analyses of the three language versions of the documents to determine whether the document groupings and projections in one language were related to the groupings and projection in another language. The projections were strongly related, among all three languages

The relationships between the vectorizations of the English, French, and Spanish versions of the documents were formally verified using a statistical test. The resulting visualizations showed a strong coherence between the document clusters in all three languages. The relationships between the document clusters can be construed as a crude mapping between the respective information spaces. These relationships have potential applications in cross-language information retrieval and document analysis.

Section 2 describes the data in more detail. Our processing of the data is described in Section 3. The results of the analysis are contained in Section 4. Implications, extensions and applications, as well as related work in the open literature, are described in Section 5.

2. Description of Test Data

The corpus used for our analysis is a subset of the United Nations Parallel Text Corpus (Version 1.0) which we obtained from the Linguistic Data Consortium (LDC). This data consists of documents in English, French and Spanish. Many of the documents appear in only one or two of the three languages. In addition, significant portions of the documents have errors in them – in terms of categorization, translation, or missing portions of the text.

To select a suitable test set, we wrote a Perl script that checked for the availability of each document in all three languages. If a given document occurred in only one or two of the three languages, it was not used in this study. In addition, the script checked for the number of lines in each version of the document. If there was a difference greater than 15 lines in length between any two of the versions of a document, that document was not used. While this is far from a foolproof method for determining that two documents are, in fact, translations of each other, it served as an acceptable first approximation. For example, the French version of a given document may have been 45 lines in length and the English version 127 lines in length – such occurrences were not uncommon. Our crude script was adequate to determine that there was a likelihood error between these two versions.

Running this Perl script against the full corpus left us with a total of 33,429 documents (11,143 in each language), including documents dated from 1988 to 1993. To reduce the disk space and CPU cycles associated with processing our test data, we used only the documents from 1988 and 1989. This provided us with a final corpus size of 9,060 documents (3,020 in each of three languages).

The documents were all tagged in SGML, and a separate Perl script was written to remove these tags before processing by SPIRE. The average final document length was 21 KB (roughly 10 pages). Most of the documents were about 6 pages, and a few were quite long. It is possible that results could be improved by using shorter documents (e.g., one or two pages per document would probably be ideal) or by automatically dividing the documents into sub-documents and processing each of them separately.

3. Methods: Document Feature Identification with SID

In order to process the data, we had to make some minor modifications to our system, enabling it to recognize non-English ASCII characters. Normally, we ignore the upper 128 positions in the

ASCII character table, to facilitate faster processing. However, the non-English documents in this corpus used the ISO-8859 character tables, in which the upper 128 character positions map non-English characters. We did not modify our stop-word list, since it was assumed that most of the words in the English version of this list would be ignored on the basis of our statistical analysis – though at the cost of a little more CPU time. It is possible that ignoring the stop-word list while processing the English documents would have given us more similarity between the respective projections we made of the different language versions of the corpus, since we did not use a stop-word list with either the French or Spanish corpora.

Several visualizations were made of the data. First, we created a visualization that combined all three language versions of the documents. Secondly, we created separate visualizations of each language version of the corpus to get a rough idea of the differences and similarities among them. Finally, we created three visualizations (one for each language) in which we endeavored to match the parameters as closely as possible to one another, in order to assure ourselves that we were "comparing apples and apples." The results we present here are based primarily on the third set of visualizations.

SID, or the System for Information Discovery, is the signal generation engine associated with SPIRE. It relies on statistical methods to determine the number of dimensions that will be used to create a visualization of the corpus. Parameters can be adjusted to indirectly select the number of dimensions each vector will contain. We targeted 150 dimensions for this comparison, although (due to the indirect determination of the number of dimensions) we ended up with 152, 151, and 154 for the English, French, and Spanish vectors, respectively. We chose to create 100 clusters for each of the three visualizations.

4. Results

First, the results from the corpus containing the English, French, and Spanish documents are described. Then the results of the separate analyses are presented.

4.1 Combined Corpus

As expected, the set containing all three languages resulted in a visualization with three discrete clouds of documents. This is because the three sets have almost no words in common with one another. This set was processed with 600 dimensions, under the assumption that a typical projection of the documents from each respective language would use between 100 and 200 dimensions. It was not determined whether or not the resulting vectors contained comparable dimensional contributions from each language or whether the dimensions were asymmetrically distributed among the three languages. This might be an interesting item to explore in follow-up work.

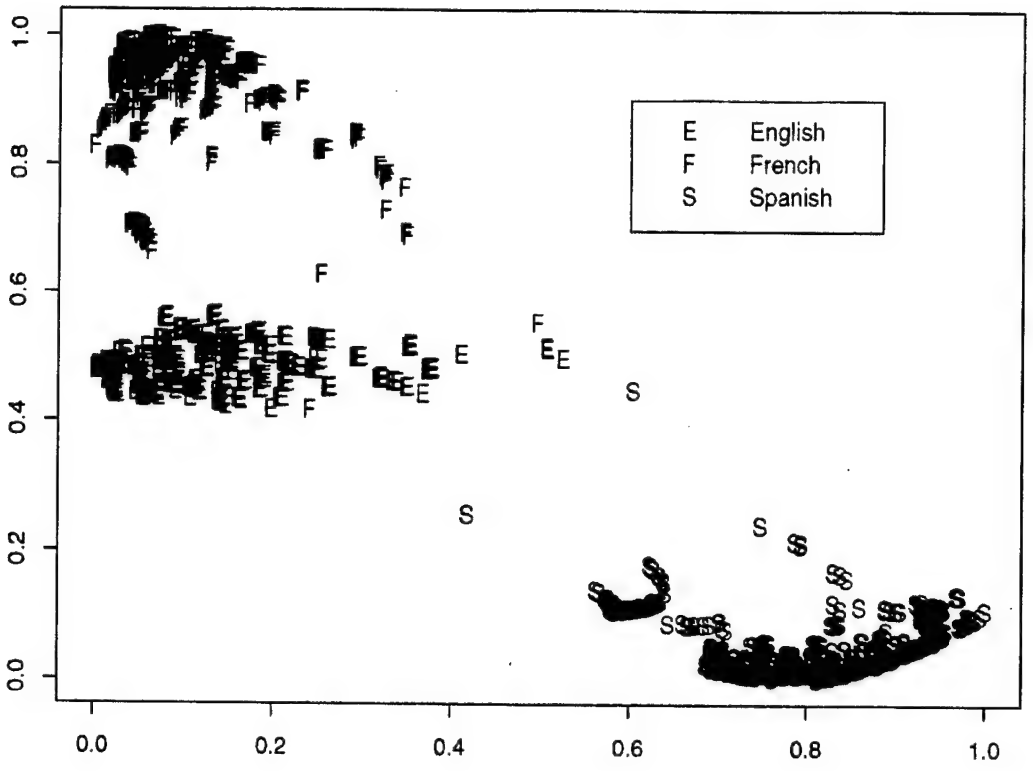


Figure 1: Galaxy display of the combined document collection. The separation between languages is somewhat expected, but not guaranteed due to the low dimension view of the data shown.

This visualization of the combined data was produced as an aid to tuning the final projections. Because default parameters were used, the numbers of dimensions and clusters varied from one language to another (even though the user can select the number of clusters, empty clusters get dropped to improve processing speed).

4.2 Separate Corpus

A separate projection was made for the document corpus in each language. The first set of plots below (Figure 2) shows the SPIRE Galaxy views of the English, French, and Spanish documents (in that order). A point represents each document. The colors were chosen based on the cluster membership in English; there are more clusters than available colors, so the same color (or gray-scale) is used for multiple clusters. A general observation is that a tight clump in one language tends to get mapped into several clumps in the others. We initially explored this relationship graphically by selecting clusters in one language and looking at the resulting points in the others. Points within a cluster were selected from the English graph, and the corresponding documents shown in the other two plots. Two samples of this selection process are shown below in Figures 3

and 4. The order of plots are the same in Figures 2, 3, and 4 (English, French, Spanish) and the plot scales are also constant across these three figures.

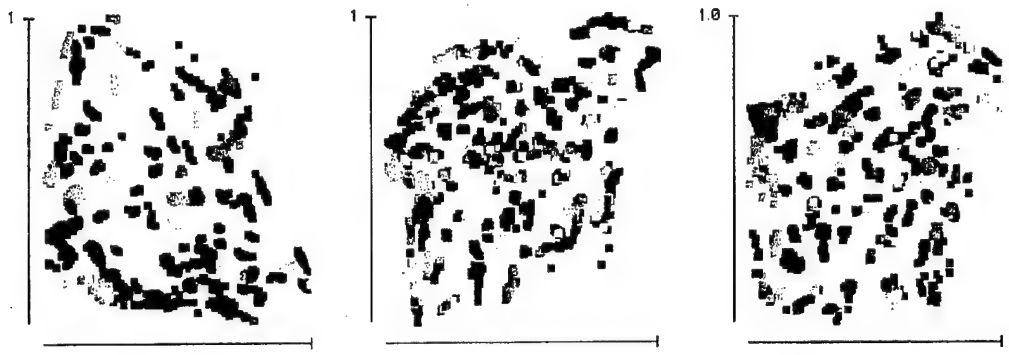


Figure 2: The English, French, and Spanish documents as projected by SPIRE. The coloring links documents across the three graphs.

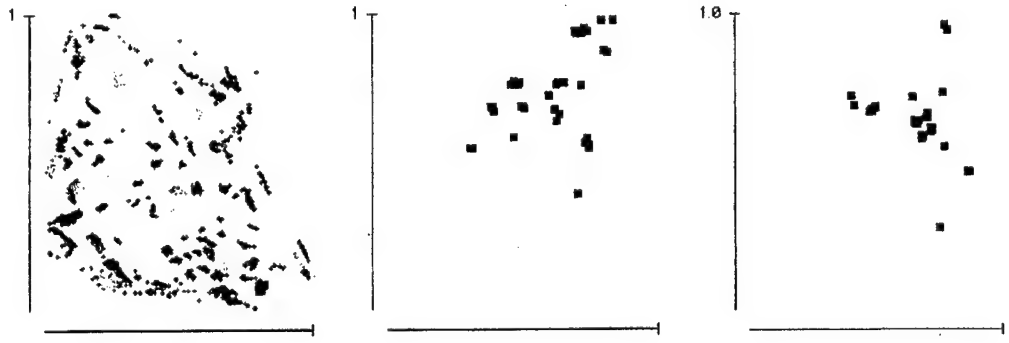


Figure 3: A cluster (lower right in the first view) selected from the English documents and the corresponding French (middle view) and Spanish (last view) documents.

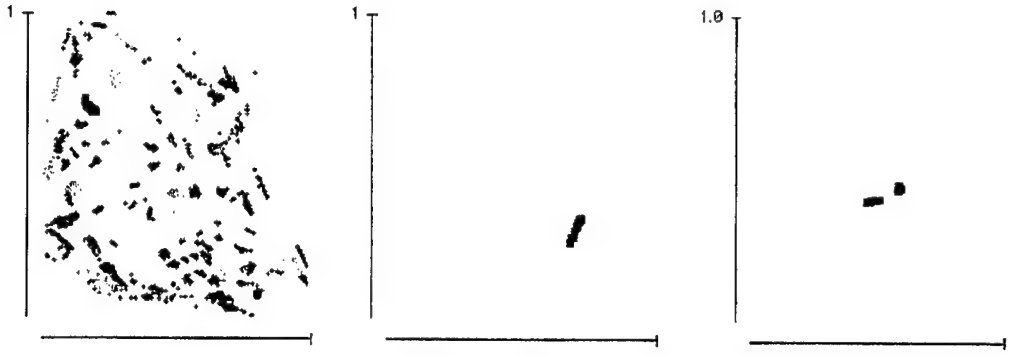


Figure 4: A cluster (lower right in the first view) selected from the English documents and the corresponding French (middle view) and Spanish (last view) documents.

These graphics are reasonable static representations of the relationships. The interactive selection process, as it is encountered in SPIRE, is more compelling. This comparison is also limited in that it shows only the two-dimensional projection of the relationships, as opposed to the full three-dimensional relationships.

To evaluate the relationships in higher dimensions, we ran a "formal statistical test" (a chi-square test). This test strongly supports the conclusion that there is a relation between the clusters in the three languages. The mean value of the chi-square quantity, if there were no relation between the clusterings, is 9801. The calculated values are very far from this nominal value; hence, we know that there is a real relation between the clusterings in the different languages. However, the chi-square value does not provide any indication of the nature of the relation.

Table 1: Comparison of clustering for three language versions. The value is compared to a chi-square distribution with 9801 degrees of freedom.

	Chi-square value 9801 df
Spanish/French	107,000
Spanish/English	99,500
English/French	98,600

To better display the relationships among the three information spaces, we built three matrices. In the first, a matrix with all the clusters from the English corpus arranged along one axis and all the clusters from the French corpus along the other, each cell in the matrix contained the num-

ber of documents that are common to both clusters represented by those X,Y coordinates. Because there is some dimensional variability between the different corpora (i.e., it is unlikely that dimension #47 in the French corpus will be equivalent to dimension #47 in the English corpus), the cluster numbers are not equivalent between the two corpora. Indeed, we can reasonably expect that cluster #47 in English might correspond with more than one cluster in the French information space. However, by identifying the cluster numbers containing the highest number of common documents between the French and English corpora, it is possible to identify the equivalent clusters between the two sets.

In order to accomplish this task, we wrote a small program to sort the cluster sequence and match the equivalent clusters between the two corpora. The criteria used by this heuristic algorithm was to permute the cluster labels in each language so as to concentrate high-count cells in either the main diagonal or one of the off-diagonals. This results in the common clusters being arranged along the dominant diagonal. We ran this program against each pair of languages, after forming the appropriate matrices in the manner described above. The resulting mappings between information spaces are shown in Figures 5, 6, and 7. The color is mapped to the number of documents in each cluster (Figures 6 and 7 appear at the end of this paper). The color code is used to indicate the number of documents in each cluster.

The clustering/clumping of the documents seen in this test-aligned corpus is somewhat language (English/French/Spanish) independent. There appears to be a very good chance of creating a useful map between the document vectors in one language and the corresponding vectors in another.

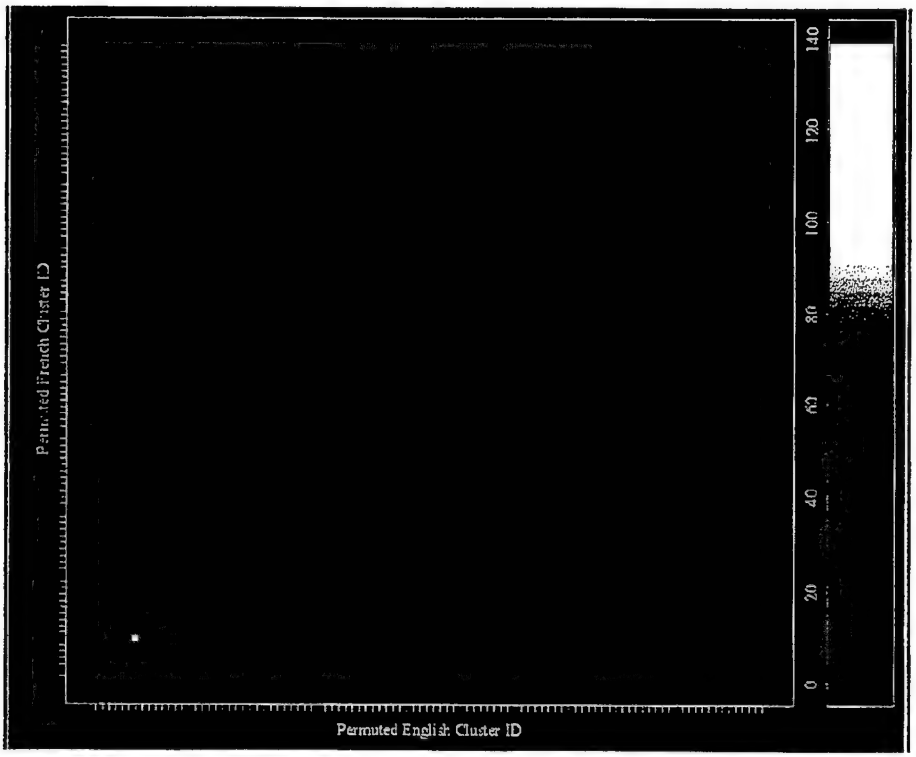


Figure 5: Comparison of cluster membership for French and English versions of the documents.

A numerical summary of the degree of "coherence" in the mapping is given in Table 2. While these numerical summaries are more intuitive than the chi-square summary presented in Table 1, they unfortunately cannot be normalized for incidental parameters like the number of clusters and the number of documents in the parallel corpora. In particular, if we set the clustering in SPIRE to 50 clusters (instead of 100) for each language, then we would expect the proportion of documents along the diagonal to differ significantly from the nominal 50% value seen in Table 2. Subject to that warning, Table 2 presents for each language pair the proportion of documents on the main diagonal in Figures 5, 6 and 7, and the proportion of documents on the main diagonal plus the adjacent off-diagonal.

Table 2: Summary of co-classification for English, French and Spanish documents.

Proportion co-classified	English/French	French/Spanish	Spanish/English
Main diagonal	0.50	0.51	0.47
Tri-diagonal	0.62	0.68	0.64

A mapping between the information spaces represented by the aligned corpora can be constructed from English to French documents as follows:

For a candidate English document:

1. Calculate the document vector (based on the existing English document set)
2. Determine which cluster (in English document space) the vector belongs to
3. Here are two options for finishing this algorithm:
   a. Map to the center of the cluster containing the majority of the corresponding French documents. Thus, the target of this mapping is a single point in "French space".
   b. Use the corresponding French documents to create a distribution within French space, return this distribution as the result of the mapping.

Note that neither of these mappings depends on being able to obtain a "nice looking" permutation of the cluster labels. Applications of these mappings are discussed in Section 5.

5. Implications and Applications

For all of the proposed applications, we assume that we have a sufficiently large parallel corpus available. All the applications envisioned hinge on the mapping between information spaces like those exhibited in Figures 5, 6, and 7. Additionally, these applications hinge on 1) being able to extend the vectorizations of the information space to new documents and 2) the parallel corpora that are the basis for the mappings being rich enough to adequately represent the new documents. In principle, item 1) is straightforward for many alternative document vectorizations (including the vectorization strategy associated with SPIRE). Item 2) is a standard requirement for many empirical strategies. It is conceivable that a useful criterion can be constructed to cover the adequacy of the existing parallel corpora for evaluating the proposed new document. Setting aside item 2, here are some applications:

5.1 Information Retrieval in Multiple Languages

The problem is to find interesting "other" language documents using your language (say, English) to form the query. Possible steps are:

1. Use the parallel corpus to establish a mapping between document vectors of the English and Spanish (for instance) documents.
2. Formulate the query in English

3. Map the English query to the Spanish information space using the mapping derived in Step 1.

4. Select documents within the neighborhood of the mapped query for translation.

5.2 Triage Documents for Translation

The problem is to find interesting "other" language documents to be translated in detail. Assuming that "interestingness" can be adequately captured as a query, this problem maps into the information retrieval problem just described.

5.3 Viewing Multiple Language Documents in the Same Space

The steps are:

1. Use the parallel corpus to establish a mapping between document vectors of the languages.

2. Choose one of the information spaces (presumably the user's language-of-choice); map the other documents into this space. If the mappings are distributions (as in Step 3b of the map construction procedure in Section 4.2), the other-language documents will be represented by a point-smear.

5.4 Related Work

An excellent summary of recent work in translingual information retrieval is provided in [CARB97]. Standard techniques mentioned therein include various limited applications of machine translation technology to queries, and taking full advantage of parallel corpora. The way in which a parallel corpus is used in previous work differs from the strategy we took. For instance, in [LIS96] the parallel corpora are used to create a single corpus by combining both French and English versions of the document into a single document. Then either French or English queries can be constructed, and either language version of the document can be retrieved.

References

[LDC94] United Nations Parallel Text Corpus (English, French, Spanish) 1994. Available from the Linguistic Data Consortium,

http://www.ldc.upenn.edu/ldc/catalog/html/text_html/unptc.html.

[LSI96] Dumais, S. T., Landauer, T. K. and Littman, M. L. (1996) "Automatic cross-linguistic information retrieval using Latent Semantic Indexing." In SIGIR'96 - Workshop on Cross-

Linguistic Information Retrieval, pp. 16-23, August 1996.
http://superbook.bellcore.com/~std/papers/SIGIR96.ps

[SPIRE95] Wise, J.A., Thomas, J.J. et al 1995. Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Documents. Proceedings of the IEEE '95 Information Visualization, IEEE Service Center, 51-58. Atlanta GA.

[CARB97] Jaime Carbonell, Yiming Yang, Robert Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. "Translingual Information Retrieval: A Comparative Evaluation". In Proceedings of Fifteenth International Joint Conference on Artificial Intelligence} (IJCAI-97).

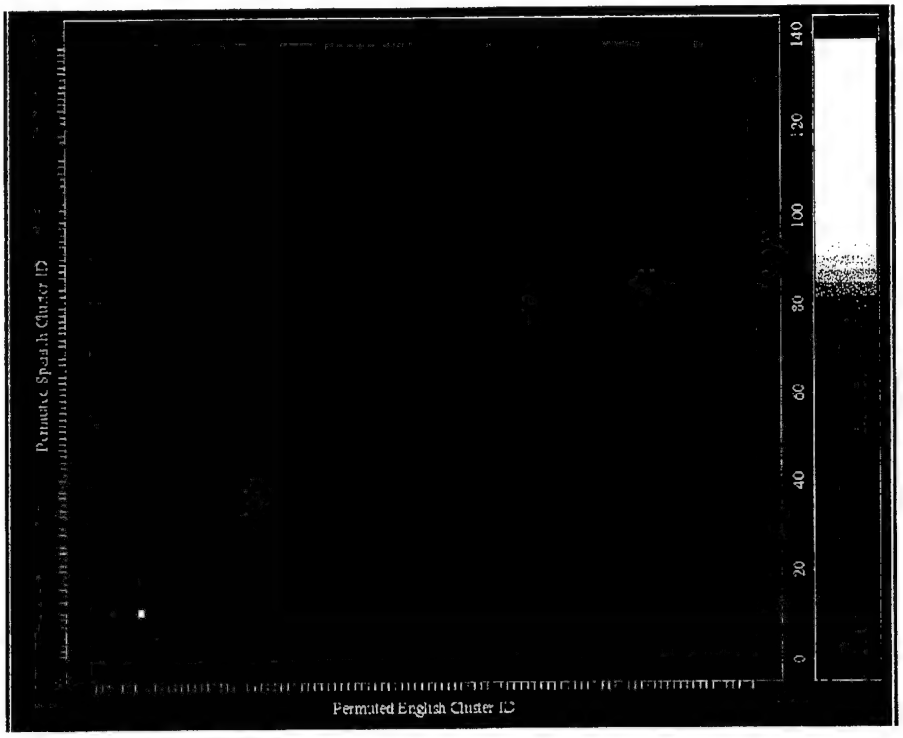


Figure 6: Comparison of cluster membership for Spanish and English versions of the documents.

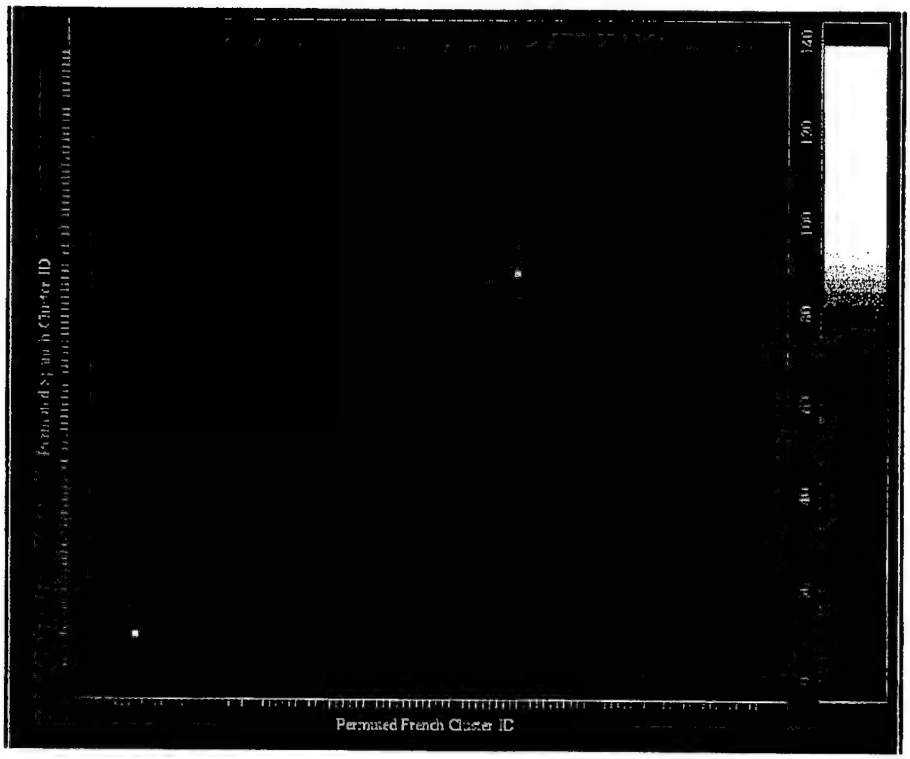


Figure 7: Comparison of cluster membership for Spanish and French versions of the documents.

APPENDIX F

REFERENCES FOR THE

LANGUAGE TO MATHEMATICS TASK

References for the Language to Mathematics Task

Paul Whitney, Battelle PNNL- Last Modified December 17, 1997

Table of Contents

People and Organizations

Professional Societies

- Association for Computational Linguistics (ACL)- An international society of individuals engaged in problems involving natural language and computation.
- The Association for Mathematics of Language (MOL)- A special interest group in the Association for Computational Linguistics MOL promotes work in mathematical linguistics.
- Special Interest Group on Information Retrieval (SIGIR) - SIGIR is a group within the Association for Computing Machinery (ACM)whose members focus on information retrieval.

Conferences

Recurring Conferences

- Text Retrieval Home Page (TREC) - A series of conferences co-sponsored by NIST and DARPA; the conferences are held to encourage research in IR. Contains pointers to the actual conference proceedings.
- ACM SIGIR Home Page - Annual Conference for the Special Interest Group for Information Retrieval. Also see: SIGIR, SIGIR Conference Page or SIGIR Information Server. Here's a specific link to one of the conferences: 18. SIGIR 1995: Seattle, Washington, USA
- MT Summit IV - A conference intended to bring together people interested in machine translation.
- http://carbon.cudenver.edu/~bstilman/firstsymp.html- The First Symposium on Linguistic Geometry and Semantic Control

Other conferences and workshops

- ANLP Workshop "Tagging Text with Lexical Semantics: Why, What, and How?"- A workshop dedicated to augmenting text with lexical information. My impression from the papers was that the "Why" part is either understood or ignored.
- A pointer to numerous linguistics conferences is maintained at http://www.emich.edu/~linguist/conference.htmlby the Linguist List

Journals

- INTELLIGENT DATA ANALYSIS is a journal focusing on application of artificial intelligence techniques to data analysis.
- Computational Linguisticsis a journal focusing on the design and analysis of natural language processing systems. The journal is associated with theAssociation for Computational Linguistics
- International Journal of Digital Librariesis a journal focusing on technologies and ideas related to digitally stored information.
- Machine Translationis a journal devoted to Computational Linguistics and related fields.
- Journal of Machine Translation
- Journal of Computational Linguistics
- Journal of Quantitative Linguistics- Publishes results on quantitative characteristics of text and language. The journal of the International Quantitative Linguistics Association (IQLA)

University Departments and Research Groups

- Institute for the Learning Sciences - A research group at Northwestern University that applies diverse tools (including natural language processing (NLP)) to build educational software.
- UTCS Neural Nets Research Group - Research group in the University of Texas at Austin Computer Science Department that focuses on "artificial intelligence and cognitive science, including natural language processing, episodic memory, schema-based vision, self-organization in the visual cortex, and neuro-evolution in sequential decision tasks such as game playing and robotics." The group is headed by Risto Miikkulainen
- TraumAID - A program to develop computer assistance to physicians for the diagnosis and treatment of penetrating trauma (e.g., gunshot wounds) and other emergency medicine.
- USC Information Sciences Institute - Research group at the University of Southern California that focuses on machine translation, AI, and related areas.
- University of Maryland Machine Translation - The objective of this group is interlingua based translation. This group is part of the Computational Linguistics and Information Processing Laboratory at the University of Maryland.
- Center for Machine Translation at Carnegie Mellon University - A large (>50) research group focusing on machine translation. Notable projects include Kant; a knowledge based machine translation system, as well as other translation and information retrieval projects (see here).
- The Language Technology Group - A technology transfer organization associated with the University of Edinburgh The technological focus is "language engineering". They have products for syntactic analysis of English text information retrieval, and hypertext document analysis.
- Centre for Cognitive Science - Also at the University of Edinburgh This group focuses on language processing and cognitive science, "often from a computational perspective". Interestingly, this centre was founded in 1969 (very early) as the School of Epistemics.
- Natural Language Processing Group at the University of Edinburgh.
- Center for Spoken Language Understanding (CSLU) - CSLU is part of the Oregon Graduate Institute of Science and Technology. The focus is spoken language understanding with an emphasis on education, technology transfer and research.
- Cognitive Computer Science Research Group - A research group at the University of California, San Diego UCSD focusing on knowledge representation, for both natural and artificial intelligence.
- Institute for Neural Computation - Home page for the UCSD Institute for Neural Computations. Various work in neural networks; diverse applications including speech processing. Collected reports in http://www.cnl.salk.edu/cgi-bin/pub-search
- The Neuroengineering Laboratory is part of the UCSD Institute for Neural Computations. The lab focuses on pattern recognition applications.
- Center for Research in Language - Yet another group at UCSD interested in language processing, cognitive science and related fields.
- Center for Language and Speech Processing - Johns Hopkins University center for creation of automatic language processing systems.
- IRCS Homepage - The home page for the Institute for Research in Cognitive Science of the University of Pennsylvania. Technical report abstracts from IRCS are found here.
- Formal Reasoning Group - The Stanford University formal reasoning group has focused on modeling context in logic systems.

Courses

- Carnegie Mellon University Spring 1995 Natural Language Processing I- The course objectives are to explain standard language transformations, morphology, and syntax, as well as to discuss phonology.
- Course syllabi- collection of syllabi from Christopher Manning (including the course listed immediately above).
- 600.465 Natural Language Processing Home Page - Home page for a Natural Language Processing course at Johns Hopkins.
- The Machine Translation short course is a three-day overview of the state of the art in computerized language translation.

Government

- Human Language Systems Program (HLS) - DARPA programs.
- Defense Advanced Research Projects Agency- Supports diverse research in areas of potential military importance. Projects relating to language processing.
- Information Access and User Interface Division- A NIST organization focusing on standards in the areas of information exchange and access; especially multimedia objects.
- U.S. Army Medical Research and Materiel Command- Overview of medical research areas.
- SAIRE Homepage - NASA project to provide to expert and novice access to meta-data related to earth and space science.

Commercial

- SRI International - A nonprofit R&D firm with some language processing research contracts.
- Language Technology Group - Commercial "arm" of University of Edinburgh's Language dept.
- Microsoft Research Natural Language Processing Group - The branch of Microsoft Research focusing on language processing.
- Natural Language Processing at Canon Research.
- The Science and Technical Research Laboratories the research arm of the Japanese Broadcasting Corporation does some Machine Translation research.

Battelle

Note that these links will work only on local Battelle computers

- P1000 Linkages and Association
- SID Text Engine Design Notes
- Pac2 lesson - Analysis with SPIRE
- http://cloak.pnl.gov/
- http://mirrors.pnl.gov/
- SID Text Engine Design Notes

Translation Companies

Links to numerous translation companies are here and here.

- AppTek - A partner with PNNL on a DARPA proposal. AppTek's expertise include

Arabic/English machine translation.
- Lingsoft - A company that has broad coverage of software for standard linguistics tasks.
- Systran - One of the oldest translation companies around. Spun off from early DARPA work. They have a demo service to translate web pages "while you wait."
- Globalink - Makers of inexpensive language translation software for PCs.
- IBM Translation Manager - Translation Technology for Windows and OS/2.

Some People

- Eduard Hovy - An instructor for the Machine Translation short course and member of USC's ISI.
- Kevin Knight - An instructor for the Machine Translation shortcourse and member of USC's ISI.
- David Yarowsky - Key work in using statistics to disambiguate words (that is, decide which meaning of a word was intended).
- David M. Magerman - Used statistics and collection of parsed sentences to create a parser.
- Risto Miikkulainen - Author of *Subsymbolic Natural Language Processing*. This book summarizes a lot of work related to understanding text descriptions of focused situations.
- Eugene Charniak - Author of *Statistical Language Learning*. Selected publications are listed in DB&LP: Eugene Charniak
- Gerard Salton - List of publications available at http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/s/Salton:Gerard.html as well as DB&LP: Gerard Salton.
- Abraham Bookstein - List of publications available at DB&LP: Abraham Bookstein
- Thomas M. Cover - Resource in information theory. Review of relevant information from text Elements of Information Theory provided below.
- G.K. Zipf - Pioneer in statistical linguistics.
- Ralf Brown - Researcher in statistics applications to language processing, especially in the areas of translation and information retrieval.

Reference and Archive Sites

Some collections of references are associated with University research groups; see the listing above.

Information Retrieval

- Some Papers on Information Retrieval - as advertised: some papers on information retrieval.
- Big Picture: Visual Browsing in Web and non-Web Databases
- Latent Semantic Indexing is a methodology for creating vector representations of documents and queries for documents. The document feature used to create the vector is the word frequency.
- http://www.dlib.org/dlib/june96/06contents.html - On-line magazine on information retrieval
- Compound Key Word Generation from Document Databases Using A Hierarchical Clustering ART Model - An article from the online journal, Intelligent Data Analysis. The article presents ideas on how to create semantic associations between words based on a text database.
- Natural Language Information Retrieval: TREC-3 Report T. Strzalkowski, J. P. Carballo, M. Marinescu (New York University) p39.
- Writings by Ralf Brown - Some excellent IR papers.

Cross language information retrieval information:

- DELOS Zurich Workshop - Cross language information retrieval workshop
- Cross-Language Information Retrieval Resources - A general resource on cross-language information retrieval.
- Cross-Language Text And Speech Retrieval - The web-site for the 1997 AAAI Spring Symposium on Cross-Language Text and Speech Retrieval.
- Experiments in Multi-Lingual Information Retrieval - A 1972 document describing an experiment in multi-lingual IR.

Computer Science Reference Sites

- Columbia Univ CS - technical report archive for Columbia University Computer Science Department.
- UTCS Neural Nets Research Group Publications
- HCRL Technical Reports - Human Cognition Research Laboratory technical report archives.
- Networked Computer Science Technical Reports Library - Indexed archive of online technical reports from numerous university computer science departments.
- Publications and Abstracts - This page contains a list of publications for University of Bonn Computer Science Department.
- Databases and Logic Programming - An archive containing reference information; some of which is related to information retrieval.

Linguistics Archive Sites

- The Computation and Language E-Print Archive - A repository for technical reports on computational linguistics and related fields.
- The LINGUIST Network - A pointer to ongoing discussions on linguistics; the format is a mailing list. Additionally, the LINGUIST serves as an organizing location for reviews and other research resources. For instance, here is a pointer to University Linguistics programs around the world.
- Resources on Statistical NLP/corpus-based computational linguistics - An archive web page of interesting pointers in computational linguistics.
- http://www.yahoo.com/Social_Science/Linguistics_and_Human_Languages/is the yahoo page pointing to far too many sites to look at regarding language processing.
- The Human-Languages Page - A hand crafted collection of links regarding language.
- Introductory and general information about lexical functional grammars is available from the lexical functional grammar web page
- The ACL NLP/CL Universe - An archive site in Natural Language Processing and Computational Linguistics maintained by theAssociation for Computational Linguistics
- Machine Translation Linksis a collection of pointers to various resources and information on translation.
- The Collection of Computer Science Bibliographies - A general collection of bibliographic information on writings in Computer Science.
- Computation and Language - This page is the interface to an archive of electronic reprints on Computation and Language. Part of a larger archive here.

Natural Language/Speech Processing Sites

- Natural Language Processing: Penman, Pangloss Projects- The USC/ISI natural language projects.
- Natural Language Processing on the Web - A nicely organized collection of pointers from Carnegie Mellon University.
-
- ftp://rtfm.mit.edu/pub/...Language_Processing_FAQ- The summary of language processing made for the usenet group news:comp.ai.nat-lang
- comp.speech WWW site - Contains the Frequently Asked Questions for the comp.speech newsgroup.
- NLP & AI Related Site - Pointers to numerous language processing sites, many of which are in Asia.
- Resources on Statistical NLP/corpus-based computational linguistics- This page contains pointers to numerous resources in NLP.

The following are pointers to particular Language/Speech processing references.

- Survey of the State of the Art in Human Language Technology. A monograph of the same name, and with similar content, is to appear this year. This particular group of researchers is focusing on speech recognition.
- An Association Based Method for Automatic Indexing with a Controlled Vocabulary- paper describing an approach to solving the document routing problem.
- Statistical Language Learning- Book summarizing applications of statistics to language processing. For a negative review, see: Review of Eugene Charniak's *Statistical Language Learning* by David M. Magerman. *Computational Linguistics*, 21(**2**). March 1995.
- Parsing as information compression by multiple alignment, unification and search By J. Gerard Wolff. SEECS Report CS-JGW-96-2.1, November 1996. Describes how parsing a sentence can be related to information compression.
- Finch, Steven, 1993. Finding Structure in Language, University of Edinburgh Ph.D. Dissertation.

Translation

- Machine Translation
- The KANT Project at CMU/CMT - The Kant project is a knowledge-based approach to machine translation. Initially the project is focusing on translating the technical/maintenance manuals for a manufacturer with a large international client base.
- http://duke.cs.brown.ed...ec/abstracts/learn.html - A Statistical Syntactic Disambiguation Program and What It Learns
- http://duke.cs.brown.ed...ec/abstracts/learn.html - Abstract for a paper on application of statistics to syntactic disambiguation.

Related Fields

Graph Theory

- 27-190 Lectures - Web page for an introductory course in discrete mathematics. The content seems typical for this type of course.
- Graph Theory Internet Resources - A few links to graph theory resources.
- G-Net Home Page - A fairly elaborate page with graph theory resource links.

Neural Nets

- Neural Network FAQ - The usenet overview of neural nets associated with the newsgroup comp.ai.neural-nets.
- The Neural Networks Research Center in Finland is a hotbed of work in self-organizing maps. Some software is available here. The information retrieval program WEBSOM originated here as well.
- Neural Networks at Pacific Northwest National Laboratory
- Directory of /pub/neural-nets - The main directory for the University of Texas neural net group
- DTI NeuroComputing Web is part of the UK's Department of Trade and Industry's neural computing program. Standard NN architectures are documented here.

AI and Knowledge Representation

- Artificial Intelligence FAQ- The usenet overview of artificial intellgence associated with the newsgroup comp.ai
- BUILDING PROBLEM SOLVERS - Blurb for an AI book on constructing inference systems.

Cognitive Sciences

- ECCS'97 Workshop - Yet another interdisciplinary workshop in the field of Cognitive Science. This workshop focuses on "context".
- Mathematical Models of Human Memory: Tutorials

Childhood Development

- Some Piaget related information is available at these web-sites: Piaget & the Gestalt-psychology and Piaget.
- Language Lectures - This page contains a summary of lectures on Language Development from an introductory Psychology course at Gettysburg College.
- Language Learning as Compression - Takes the thesis that ideas from information theory can be used to describe much of how language is learned. A related article is stored here on the Internet.

Can Chimps Talk?

This question is, apparently, very controversial. The following two links provide some data related to the controversy.

- TRANSCRIPT OF THE NOVA PROGRAM "Can Chimps Talk?"
- Language in child and chimp?- A collection of links about language in primates.

- Language and the Primate Brain Martin I. Sereno Cognitive Science Department, UCSD.- An interesting article on the physiological basis of language.

Information Theory

- Entropy on the World Wide Web - A collection of pointers to information theory related documents on the world Wide Web.
- 15-850(B): Information Theory - A home page for an information theory course.
- Cover, Thomas M. and Joy A. Thomas, (1991). *Elements of Information Theory*. John Wiley and Sons.

Data Sources

- For this problem, much of the data are text. The web is a source for text. The following links point to multi-lingual text sources (Biblical) and various transcribed literature.
- gopher://ftp.std.com:70/11/obi/book/ - Sources of on-online literature (e.g., the Federalist papers, Grimm's Fairy Tales, ...)
- The Electronic Text Center at the University of Virginiacontains numerous texts.
- Concordances and Corpora Tutorial - This page is the entry to an online course on gathering text data. Catherine N. Ball of the Georgetown University Department of Linguistics created the course.
- Corpus Linguistics - Information related to collections of text (includingparallel corpora) available for study, testing etc.
- Multi-Language Editions - Source for multilingual bibles.
- Linguistic Data Consortium - A group of various research organizations that collect and distribute speech and text databases and other R&D resources. Also the source of the parallel corpus used by Battelle in the report *Applications of Parallel Corpora in Document Analysis* by PD Whitney and DL McQuerry.
- Electronic Text Center at the University of Virginia- pointers to English, French, German, Japanese, and Latin texts.

Various translations and versions of the bible are the subjects of this next block of web pages; these are good sources of matched sentences and phrases in diverse languages.

- The Bible, Revised Standard Edition, at the Electronic Text Center, University of Virginia
- Virtual Christianity: Bibles

Software

- SMART - SMART is an implementation of the vector-space model of information retrieval proposed by Salton back in the 60's. The primary purpose of SMART is to provide a framework in which to conduct information retrieval research. Standard versions of indexing, retrieval, and evaluation are provided. The system is designed for use with small to medium scale collections, and offers reasonable speed and support for these actual applications. SMART analyzes the collection of information and builds indexes. It can then be used to build natural language based information retrieval software. It uses feedback from the user to tighten its search. Also see Directory of /pub/smart

- Inquery - Information retrieval code based on Bayesian ideas.
- PC-PATR is a syntactic parser available for anyone's use.
- LINK: A Combinatorics and Graph Theory Workbench for Applications and Research- **LINK** is a set of C++ libraries that supports applications in discrete mathematics.
- Text Analysis Software - This page contains utilities for text analysis available at the Electronic Text Center at the University of Virginia
- The Summer Institute for Linguistics has a software archive here.